Journal de la Société Française de Statistique
Vol. 151 No. 1 (2010)

Model evaluation in nonlinear mixed effect models, with applications to pharmacokinetics

Titre: Evaluation des modèles non-linéaires à effets mixtes, avec une application en pharmacocinétique

Emmanuelle Comets [1], Karl Brendel [2] and France Mentré [1]

Abstract: Model evaluation is an important part of model building, and has been the subject of regulatory guidelines in drug development. In the present paper, we illustrate the use of some recently proposed metrics on several simulated datasets. These metrics include Visual Predictive Checks (VPC), prediction discrepancies (pd) and normalised prediction distribution errors (npde). We illustrate them using simulated datasets. Prediction bands around selected percentiles can be obtained through repeated simulations under the model being tested, and their addition to VPC plots or plots of pd and npde versus time and predictions are useful to highlight model deficiencies. Tests for some of the metrics are also available and can be used as a complement to graphs.

Résumé : L'évaluation est une partie importante de la construction de modèles, faisant l'objet de recommendations de la part des autorités régulant la mise sur le marché de nouveaux médicaments. Dans ce papier, nous effectuons une courte revue de métriques récemment proposées, en particulier les VPC (Visual Predictive Check), les discordances de prédictions (pd) et les erreurs de prédiction sur la distribution (npde). Nous illustrons ces métriques sur quelques exemples simulés. Nous montrons comment il est possible de construire des bandes de prédiction autour de la courbe des médianes (ou d'autres percentiles) des données simulées. Ces bandes de prédiction sont un outil visuel particulièrement efficace pour détecter des zones où le modèle peut être amélioré. La distribution de certaines métriques est connue et permet de proposer des tests pour compléter les graphes diagnostiques.

Keywords: Nonlinear mixed effect models, model evaluation, VPC, npde
Mots-clés : Modèles non-linéaires à effets mixtes, évaluation de modèles, VPC, npde
AMS 2000 subject classifications: 92B15, General biostatistics

1. Introduction

As recently defined by Ette and Williams in the eponym book, pharmacometrics is the science of interpreting and describing pharmacology in a quantitative fashion [13], and can thus be understood as the quantitative science of drug development. It consists in modelling the data collected in drug clinical trials, developing and applying statistical methods to characterise, understand and predict drug behaviour, as well as in characterising the uncertainty associated with these different elements, to guide rational decision-making. Clinical trials are becoming more complex, routinely now collecting longitudinal data. Nonlinear mixed-effect models, also termed population analyses, are therefore increasingly used, in order to represent complex nonlinear processes and to describe both between and within subject variability. Evaluation is an important

[1] INSERM UMR738, Université Paris Diderot, France
E-mail: emmanuelle.comets@inserm.fr and E-mail: france.mentre@inserm.fr
[2] IRIS, Servier, France
E-mail: karl.brendel@fr.netgrs.com

 ISSN: 2102-6238

part of modelling, and a section on model evaluation has been included in the guidelines on population analyses issued by the main drug regulatory agencies [14, 11]. Evaluation methods have been the topic of a number of publications in particular in Bayesian litterature.

Over the years, different terms have been used, including *qualification, adequacy, assessment, validation, checking, appropriateness* or *performance*. The FDA guideline, issued in 1999, mentions model *validation* while the more recent EMEA guideline devotes a section to model *evaluation*. This evolution recognises the insight in Box's famous quote "all models are wrong, but some are useful", which indeed implies that no model can ever be accepted, but only evaluated with respect to key features and model use [5]. Model evaluation therefore covers notions of adjustment (whether the model describes observed data properly), parcimony (whether the model is simple enough to ensure extrapolability) and predictive performance (whether the model can be used for the purpose it has been developed for).

However, a recent bibliographic review of all population pharmacokinetic and/or pharmacodynamic models published over a four years period (2001-2004) showed that evaluation was absent from the report in 30% of the 478 models identified in the 324 papers included in this review [9]. In addition, only in 25% of the papers was the evaluation judged to be good or excellent. This unfortunate situation is partly due to the lack of a gold standard for model evaluation, but modellers do have an array of tools at their disposal. It is the objective of the present paper to provide a brief overview of these methods. We will focus on methods proposed for evaluation of models built to describe the evolution of continuous responses. This paper follows up and relies on papers by Mentré and Escolano [33], Brendel et al. [7, 8] and Comets et al. [10].

2. Models and Methods

Let B denote a building (or learning) dataset and V a validation dataset. B is used to build a population model called M_B. Evaluation consists in comparing the predictions obtained by M_B, using the design of V, to the observations in V. V can be the learning dataset B (internal evaluation) or a different dataset (external evaluation). The null hypothesis (H_0) is that data in the validation dataset V can be described by model M_B.

2.1. Statistical models

Continuous longitudinal data can be characterised by the following relationship between the observations y_{ij} collected for a value x_{ij} of the design variables (usually the time and doses in PK/PD studies) in subject i (where i=1,...,N) with covariates $\mathbf{z}_i$, and the vector of parameters θ_i characterising individual i:

$$y_{ij} = f(\theta_i, x_{ij}, \mathbf{z}_i) + g(\theta_i, \gamma, x_{ij}, \mathbf{z}_i)\varepsilon_{ij} \tag{1}$$

f is the structural model, describing the evolution of the process being modelled, and g characterises the measurement error model, which can depend on structural model predictions and on additional variance parameters γ. ε_{ij} is assumed to be normally distributed with mean 0 and unit variance. A commonly used model is the combined error model, where the variance depends of f and on additional parameters a, b and c, with c often fixed to 1:

$$g(\theta_i, x_{ij}, \mathbf{z}_i) = a + b\, f^c(\theta_i, x_{ij}, \mathbf{z}_i) \tag{2}$$

 ISSN: 2102-6238

In the following, we will denote with a bold font the vector of individual observations for subject i.

The second level of variability characterises variations between the different individuals. Interindividual variability (IIV) is usually modelled parametrically, with the vector of individual parameters θ_i modelled as a function of a vector of fixed effects μ, of individual covariates $\mathbf{z}_i$, and a P-vector of individual random effects η_i, through a function $h(\mu, \mathbf{z}_i, \eta_i)$. A common choice for the η_i is to assume they follow a multivariate normal distribution with variance-covariance matrix Ω:

$$\eta_i = (\eta_{i(1)}, \eta_{i(2)}, ..., \eta_{i(P)})^T \sim \mathcal{N}(0, \Omega) \tag{3}$$

An often used transformation h is the following, leading to a log-normal distribution for the parameters:

$$\theta_{i(p)} = \mu_{(p)}\, e^{\eta_{i(p)}} \tag{4}$$

In these equations, the subscript $_{(p)}$ denotes the p$^{\text{th}}$ component of the corresponding vector. Other distributional assumptions can be made for η, in particular an alternative approach is to consider a non-parametric distribution for the random effects [31]. Covariate relationships can be included in equation (4) through a model relating the value of μ to individual covariates, which could then be denoted as $\mu(\mathbf{z}_i)$. The dimension of the vector η may be different from the number of fixed parameters of the model, especially when covariate relationships are included, or when the interindividual variability on some parameters is assumed to be negligible.

In the following, model M_B regroups the structural model f, the residual error model g, as well as the set of population parameters Ψ. Ψ includes fixed parameters, parameters characterising the distributional assumptions on θ_i or the variance model. For instance, with a log-normal model to account for between-subject variability and a combined error model for the residual error, the vector of population parameters Ψ is $(\mu, vec(\Omega), a, b, c)^T$. Ψ is usually estimated as $\hat{\Psi}$ using the data in B.

The likelihood in nonlinear mixed effect models has no analytical expression, because it implies an integral over the distribution of individual parameters:

$$\text{ł}(\Psi\,;\,\mathbf{y}) = \prod_{i=1}^{N} \text{ł}(\Psi\,;\,\mathbf{y_i}) = \prod_i p(\mathbf{y_i}|\Psi) = \prod_i \int_{\mathscr{D}} p(\mathbf{y_i}|\eta_i, \Psi) p(\eta_i|\Psi) d\eta_i \tag{5}$$

Several estimation methods have been proposed to obtain estimates of Ψ. The first methods relied on linearisation of the model [30] or of the likelihood [42], and have been implemented in software like NONMEM and SAS. Alternatively, numeric integration can be used to obtain exact approximations of the likelihood in (5) [35], and adaptive Gauss-Hermite quadrature is the default method in SAS proc NLMIXED. Recently stochastic EM algorithms have proved extremely effective to maximise (5), and have been implemented in particular in Monolix [26].

2.2. Metrics for model evaluation

Standardised prediction errors

Prediction errors are defined as the difference between the observations $\mathbf{y}_i$ and the predictions $f(\hat{\mu}, \mathbf{x}_i, \mathbf{z}_i)$ obtained using M_B. When the residual error model is not homoscedastic, or when

 ISSN: 2102-6238

several observations are collected in some or all subjects, the prediction errors are correlated. Standardised prediction errors are obtained after decorrelation by the variance-covariance matrix of the vector of predictions, $\hat{\mathbf{V}}_i$:

$$\mathbf{spe}_i = \hat{\mathbf{V}}_i^{-1/2}(\mathbf{y}_i - \mathbf{E}(f(\theta, \mathbf{x}, \mathbf{z}))) \tag{6}$$

The matrix $\hat{\mathbf{V}}_i$ depends on the individual i through the design matrix and possibly the covariate model. The average prediction $\mathbf{E}(f(\theta, \mathbf{x}, \mathbf{z}))$ is usually replaced by $f(\hat{\mu}, \mathbf{x}_i, \mathbf{z}_i)$ assuming a first-order approximation for the model; this is the usual approach when the estimation method is also based on model linearisations, but this approximation can be poor when the model is non-linear. Alternatively, both $\hat{\mathbf{V}}_i$ and the vector of predictions $\mathbf{E}(f(\theta, \mathbf{x}, \mathbf{z}))$ can be obtained through Monte-Carlo simulations by repeatedly sampling from the distribution of the random effects. Standardised prediction errors are also referred to as weighted residuals when computed on the data used for the estimation of the parameters.

The distribution of standardised prediction errors can be determined if the model f is linear. In that case $\mathbf{spe}_i$ follow a standard normal distribution. However, when the model f is nonlinear, its distribution is no longer known because of the approximation involved in the definition of $\mathbf{spe}_i$.

Prediction distribution errors

Prediction errors appropriate for nonlinear mixed effect models have been proposed by Mentré and Escolano [33]. Let $p_i(y|\Psi)$ be the predictive distribution of an observation y given the population parameters Ψ. As the likelihood in equation 5, this predictive distribution can be obtained as an integral:

$$p_i(y) = \int p(y|\theta_i, \Psi) p(\theta_i|\Psi) d\theta_i \tag{7}$$

Let F_{ij} denote the cumulative distribution function (cdf) of the predictive distribution of Y_{ij} under model M^B, which is the integral up to the observation y_{ij} of $p_i(y|\Psi)$. We define the prediction discrepancy pd_{ij} as the value of F_{ij} at observation y_{ij}, $F_{ij}(y_{ij})$. Since the integral above is analytically intractable, F_{ij} can be computed using Monte-Carlo simulations [33]. Using the design of the validation dataset V, we simulate under model M^B K datasets $\mathrm{V}^{sim(k)}$ (k=1,...,K). Let $\mathbf{y}_i^{sim(k)}$ denote the vector of simulated observations for the i^{th} subject in the k^{th} simulation. Note that only the estimated value of the parameters is used for the simulation, neglecting the uncertainty in these estimates (known as the *plug-in* approach).

The prediction discrepancy for an observation y_{ij}, pd_{ij}, is then computed as the percentile of y_{ij} in the empirical distribution of the $y_{ij}^{sim(k)}$:

$$\mathrm{pd}_{ij} = F_{ij}(y_{ij}) \approx \frac{1}{K} \sum_{k=1}^{K} \delta_{ijk} \tag{8}$$

where $\delta_{ijk} = 1$ if $y_{ij}^{sim(k)} < y_{ij}$ and 0 otherwise.

By construction, prediction discrepancies (pd) are expected to follow the uniform distribution $\mathscr{U}(0,1)$. We can also transform pd back to a normal distribution using the inverse function of the normal cumulative density function implemented in most software:

$$\mathrm{npd}_{ij} = \Phi^{-1}(\mathrm{pd}_{ij}) \tag{9}$$

 ISSN: 2102-6238

In the following npd will be the normalised prediction discrepancies.

However, within-subject correlations are introduced when multiple observations are available for each subject. A test comparing the distribution of pd_{ij} to $\mathscr{U}(0,1)$ therefore has an increased type I error if these correlations are neglected [33]. To decorrelate the pd, we compute the empirical mean $\hat{\mathbf{E}}(\mathbf{y}_i)$ and empirical variance-covariance matrix $\text{var}(\mathbf{y}_i)$ over the K simulations. The empirical mean is obtained as:

$$\hat{\mathbf{E}}(\mathbf{y}_i) = \frac{1}{K}\sum_{i=1}^{K} \mathbf{y}_i^{sim(k)}$$

and the empirical variance is:

$$\hat{\mathbf{V}}_i = \text{var}(\mathbf{y}_i) = \frac{1}{K-1}\sum_{i=1}^{K} (\mathbf{y}_i^{sim\,(k)} - \hat{\mathbf{E}}(\mathbf{y}_i))(\mathbf{y}_i^{sim\,(k)} - \hat{\mathbf{E}}(\mathbf{y}_i))'$$

Decorrelation is performed simultaneously for simulated data:

$$\mathbf{y}_i^{sim(k)*} = \hat{\mathbf{V}}_i^{-1/2}(\mathbf{y}_i^{sim(k)} - \hat{\mathbf{E}}(\mathbf{y}_i)) \tag{10}$$

and for observed data:

$$\mathbf{y}_i^{*} = \hat{\mathbf{V}}_i^{-1/2}(\mathbf{y}_i - \hat{\mathbf{E}}(\mathbf{y}_i)) \tag{11}$$

Decorrelated pd are then obtained using the same formula as in (8) but with the decorrelated data, and we call the resulting variables prediction distribution errors (pde):

$$\text{pde}_{ij} = F_{ij}^{*}(y_{ij}^{*}) \approx \frac{1}{K}\sum_{k=1}^{K} \delta_{ijk}^{*} \tag{12}$$

where $\delta_{ijk}^{*} = 1$ if $y_{ij}^{sim(k)*} < y_{ij}^{*}$ and 0 otherwise.

Under H_0, if K is large enough, the distribution of the prediction distribution errors should follow $\mathscr{U}(0,1)$ by construction of the cdf. Normalised prediction distribution errors (npde) can then be obtained as previously:

$$\text{npde}_{ij} = \Phi^{-1}(\text{pde}_{ij}) \tag{13}$$

By construction, if H_0 is true, npde follow the $\mathscr{N}(0,1)$ distribution and are uncorrelated within an individual. The only approximation involved is to assume that decorrelation through equations 10 and 11 renders the pde independent, which is strictly true only for Gaussian variables.

Numerical and Visual predictive checks

A number of simulation-based approaches have been regrouped under the denomination of Posterior Predictive Check (PPC). These methods consider that if M_B is true, data simulated under the model should resemble the observed data. The idea is then to choose one or several statistics and compare its value when computed on V to its distribution obtained in simulated datasets through a Neyman-Pearson test. If the value is too extreme, the model is rejected. The statistic(s) can be chosen by the modeller to correspond to important model features, but it is recommended

 ISSN: 2102-6238

to choose non-sufficient statistics, that are not automatically adjusted during model fitting [2]. As an example, Girard et al. assessed model adequacy through the predictions of compliance patterns in a PK model [18]. Yano et al. studied a number of statistics for PPC, albeit in a very simple simulation setting [43].

In fact, pd and npde are a form of observation-based PPC [33]. Another very useful diagnostic tool based on PPC is the Visual Predictive Check [21]. A VPC plot is obtained by simulating a large number of datasets with the same design as V, and by plotting the prediction interval corresponding to a given value. For instance, the 90% prediction interval is obtained for each time point as the interval in which lie 90% of the simulated values for this time point. The observed data are plotted on the graph to assess whether the model is able to reproduce the evolution in time and the variability. The same simulations as those performed for npde can be used to obtain any desired prediction interval.

VPC as suggested by its name provides primarily a visual diagnostic, but tests using the simulations performed to obtain the plots have been proposed for a less subjective interpretation. Wilkins et al. computed the percentages of outliers outside several prediction intervals to the theoretical value and showed that trends in the prediction intervals can be used to pinpoint some model deficiencies [41]. This approach, which we will call PI-NPC in the following, is in fact an example of what Gelman called Numerical Predictive Check (NPC) in Bayesian analyses [16].

2.3. Tests and graphs

Tests

Under the null hypothesis that model M_B describes adequately the data in the validation dataset, npde follow the $\mathcal{N}(0,1)$ distribution. Under the additional hypothesis that the model is linear, spe follows the same distribution, although with nonlinear models the distance to the theoretical distribution can be very important even for simple PK models [33, 8]. To compare distributions, omnibus tests such as the Kolmogorov-Smirnov test can be used, or a combination of three tests: Brendel et al. proposed to use a Wilcoxon test comparing the mean to 0, a Fisher test comparing the variance to 1, and a Shapiro-Wilks test to test normality, and a global p-value can then be obtained as the maximum of the three p-values after a Bonferroni correction to account for multiple tests [7].

For PI-NPC, comparing the proportion of outliers outside of a given prediction interval to the expected proportion can be performed through a normal (approximation) or a binomial test. Here we will show the results for the 90% prediction interval, in which we will compare the proportion of points inside the interval to the value of 0.9. As with prediction discrepancies however, the type I error of PI-NPC increases with several observations per subject; Brendel et al. proposed to use instead the decorrelated observations and simulations to obtain decorrelated PI-NPC, $\text{PI-NPC}_{\text{dec}}$, for which the same tests can be used [8].

Graphs

Goodness-of-fit graphs are now widely used to examine and detect model deficiencies, and as such they are plotted for each model in the model building process. Plots of spe against time and

 ISSN: 2102-6238

predictions in particular have long been used to detect model deficiencies. The same graphs can be performed with pd and npde, which are residuals more appropriate for nonlinear models. In the field of population PK/PD, graphs of residuals versus predictions use the values predicted by the model even when the residuals have been decorrelated, as is the case for both spe and npde here. Following the suggestion of a referee, we could question whether in this case it would not be more appropriate to decorrelate also the predictions. The decorrelated predictions can be obtained as:

$$\mathbf{ypred}_{dec,i} = V_i^{-1/2}\hat{\mathbf{E}}(\mathbf{y}_i) \tag{14}$$

Comparing metrics to their theoretical distributions can be done through QQ-plots or histograms [10]. Examples of these different graphs will be shown in the next section.

VPC plots usually show the limits of the 90% or 95% prediction interval (eg, for the 95% prediction interval, the 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ percentile), sometimes shading the area lying between these two lines, as well as the predicted median; the observed data can be plotted over the interval. A very visually appealing addition to VPC plots has recently gained popularity: in addition to the 2.5$^{\text{th}}$, 50$^{\text{th}}$ and 97.5$^{\text{th}}$ percentiles, we plot the prediction interval on these percentiles. For a given time-point, we compute the values of the 2.5$^{\text{th}}$, 50$^{\text{th}}$ and 97.5$^{\text{th}}$ percentiles for each of the K simulated datasets; for each of the three percentiles, we plot the limits of the 95% interval over these K intervals. Usually the 95% intervals around each border and around the median are plotted as coloured areas; for model assessment, the corresponding 2.5$^{\text{th}}$, 50$^{\text{th}}$ and 97.5$^{\text{th}}$ percentiles of the observed data are plotted as lines or points, and should remain within the coloured areas. The same idea can be applied to graphs of pd and npde, and again examples will be shown in the next section.

3. Illustrative examples

3.1. Simulated datasets

As in [10], to illustrate the different graphs and tests, we use simulated data based on the well known toy dataset recording the pharmacokinetics of the anti-asthmatic drug theophylline. The data were collected by Upton in 12 subjects given a single oral dose of theophylline who then contributed 11 blood samples over a period of 25 hours [4]. We removed the data at time zero from the dataset, and applied a one-compartment model with first-order absorption and elimination, as previously proposed [12]. The model was parameterised in absorption rate constant k_a (hr^{-1}), volume of distribution V (L) and elimination rate constant k (hr^{-1}) and did not include covariates. V is in fact the apparent volume of distribution and should be denoted V/F where F is the bioavailability, but for the sake of simplicity we will drop the reference to F. The concentration at time t following a dose D is then obtained using the following equation:

$$C(t) = \frac{D}{V}\frac{k_a}{k_a - k}\left(e^{-k\,t} - e^{-k_a\,t}\right) \tag{15}$$

The residual variability was modelled using a combined error model. Interindividual variability was modelled using an exponential model for the three PK parameters, eg for V:

$$V_i = V\,e^{\eta_V} \tag{16}$$

 ISSN: 2102-6238

The variance-covariance matrix was denoted Ω:

$$(\eta_{Vi}, \eta_{ki}, \eta_{k_a i}) \sim \mathcal{N}(0, \Omega) \tag{17}$$

A correlation between the parameters k and V was assumed ($\text{cor}(\eta_k, \eta_V)$). Using NONMEM (version 5.1) with the FOCE INTERACTION estimation method, we obtained the parameter estimates reported in Table 1. This model and these parameter estimates correspond to M_B. Were this analysis performed in actual conditions, the large correlation between the random effects corresponding to k and V would probably be investigated, through reparameterisation and study of the relationship with covariates.

TABLE 1. *Parameter estimates for the theophylline concentration dataset. A one-compartment model was used, parameterised with the absorption rate constant k_a, the volume of distribution V, and the elimination rate constant k. A correlation between* $\ln(V)$ *and* $\ln(k)$ *($cor(\eta_k, \eta_V)$) was estimated along with the standard deviations of the three log-parameters. The model for the variance of the residual error was a combined error model.*

Fixed effects		Interindividual variability (SD)	
k_a (hr^{-1})	1.51	ω_{k_a} (-)	0.67
V (L)	31.9	ω_V (-)	0.12
k (hr^{-1})	0.087	ω_k (-)	0.13
a ($mg.L^{-1}$)	0.088	$\text{cor}(\eta_k, \eta_V)$ (-)	0.99
b (-)	0.26		

For the purpose of illustrating the different graphs and tests, we simulated datasets including N=100 subjects; in the original study, each subject had different doses and sampling times, so to design the simulation study we took the median total dose of 320 mg and the following time points, which appeared to be the nominal times in the original study: 15 and 30 min, 1, 2, 4, 5, 7, 9, 12 and 24 h. We then simulated several external validation datasets V with this design: V_{true} was simulated under M_B (H_0), using the parameters reported in Table 1. Two datasets, V_{bioavail} and V_{IIV}, were simulated assuming respectively a bioavailability divided by 2 (so that V/F is multiplied by 2), or an IIV increased by 50% for V; these situations could occurr during model development when the validation dataset is taken from a different population (eg. healthy volunteers versus patients). We also simulated a dataset with a two-compartment model, $V_{2\text{cpt}}$, with the following parameter values, to assess the influence of model misspecification: k_a=1.55 hr^{-1}, V=20 L, k=0.02 hr^{-1}, k_{12}=0.2 hr^{-1} and k_{12}=0.01 hr^{-1}, and 30% IIV on k_{12} and k_{12}. The residual error model and the values of the other interindividual variabilities were unchanged.

Figure 1 show plots of the concentration versus time profiles for the 4 datasets.

3.2. Model evaluation for simulated datasets

For each dataset, K=1000 simulations under M_B were performed. The simulations were used to plot the VPC, compute the simulation-based metrics pd and npde, and to evaluate the percentage of observed data within the 90% prediction interval for PI-NPC. Thus, in the present paper, for each setting, we simulated both the original and the external validation dataset, with different parameter sets.

 ISSN: 2102-6238

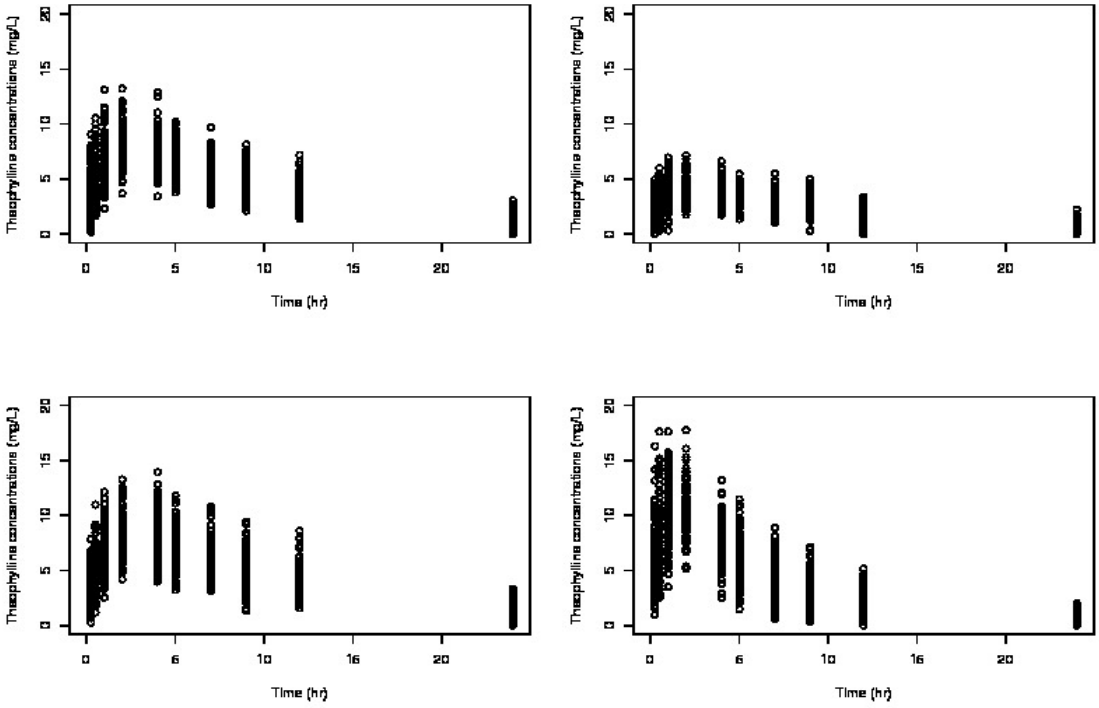

FIGURE 1. *Plots of concentrations versus time in simulated datasets* V_{true} *(upper left),* V_{bioavail} *(upper right),* V_{IIV} *(lower left),* $V_{2\text{cpt}}$ *(lower right).*

For $V_{2\text{cpt}}$, we first estimated the parameters assuming a one-compartment model, and the metrics were then computed using these new estimates instead of those from M_B. Indeed in this case we were interested in assessing whether the different metrics were able to detect the misspecification in the structural model, to mimick the use of diagnostic tools during model building where different structural models are successively tested.

3.3. Results

Tests

Table 2 gives the results of the three tests on mean, variance and distributional shape performed on the npde computed for each dataset, as well as the result of the global test obtained by combining these three tests with a Bonferroni correction [8]. We also show the alternative to the global test, a Kolmogorov-Smirnov test comparing the distribution of the npde to $\mathcal{N}(0,1)$, and the binomial test for the 90% PI-NPC. As expected, none of the tests is significant for V_{true}, but for V_{bioavail}, the mean differs very significantly from 0, and the normality assumption is clearly violated, resulting in a very significant global test. For V_{IIV}, where the variability on parameter V was

 ISSN: 2102-6238

increased from about 12% to about 25%, the variance is increased, which is detected using the Bonferroni-corrected test but not with a Kolmogorov-Smirnov test. Finally, with $V_{2\text{cpt}}$, the data simulated under the two-compartment model were fitted to a one-compartment model, but model misfit can be detected through the test and appears in the value of the test of the mean.

TABLE 2. *Values of the tests on* npde *and of the binomial test on the coverage of the PI-NPC (90% PI), for the four datasets simulated in the present study. The simulations used to compute the metrics were performed under model M_B for the first three datasets. For the data in $V_{2\text{cpt}}$, simulated under a two-compartment model, we re-estimated the parameters assuming a one-compartment model and we used these estimates to compute the metrics.*

Dataset	Separate tests (npde)			Global tests (npde)		PI-NPC
	Mean	Variance	Normality	3 tests combined	KS test	80% PI
V_{true}	0.23	0.71	0.57	0.69	0.46	0.53
V_{bioavail}	$<10^{-9}$	0.002	$<10^{-10}$	$<10^{-10}$	$2<10^{-16}$	$2<10^{-16}$
V_{IIV}	0.78	0.01	0.69	0.04	0.51	4.10^{-6}
$V_{2\text{cpt}}$	0.001	0.79	0.64	0.002	0.005	0.11

The proportion of observed data within the 90% prediction interval computed on the simulated datasets was estimated to be 89.4% (CI 87.3-91.2) for V_{true}, 38.6% (35.5-41.7) for V_{bioavail}, 85.4% (83.0-87.5) for V_{IIV}, and 88.5% (86.3-90.4) for $V_{2\text{cpt}}$. The p-values of the binomial test for the corresponding PI-NPC are shown in the last column of table 2, and show that the test was able to detect model misspecification for V_{bioavail} and V_{IIV}, but not for $V_{2\text{cpt}}$ in this example. We also performed the test on the decorrelated data: in the present example, there was hardly any difference between the p-values obtained for the PI-NPC and PI-NPC$_{\text{dec}}$, so we do not report the p-values for the latter in the table.

Graphs

The npde library also provides plots, shown in figure 2 for V_{true}. The two upper plots compare the distribution of the npde to the theoretical $\mathcal{N}(0,1)$, either through a QQ-plot (left) or with a histogram (right). The two lower plots are scatterplots of npde versus time (left) and predicted concentrations (right). For an observation y_{ij}, the predicted concentration is obtained as the empirical mean of the corresponding simulations y_{ij}^{sim} (as in the computation of $\hat{\mathbf{E}}(\mathbf{y}_i)$ for pd) (that is $E(f(\theta_i, \mathbf{x}_i, \mathbf{z}_i))$). Dashed lines at ± 1.96 indicate the interval in which we expect 95% of the npde to be, and the line in the middle indicates $y = 0$. For dataset V_{true} which is expected to correspond to M_B, no trend can be seen in the scatterplots, and the distributional plots (top) show a very good match between theoretical and observed distribution of the npde.

 ISSN: 2102-6238

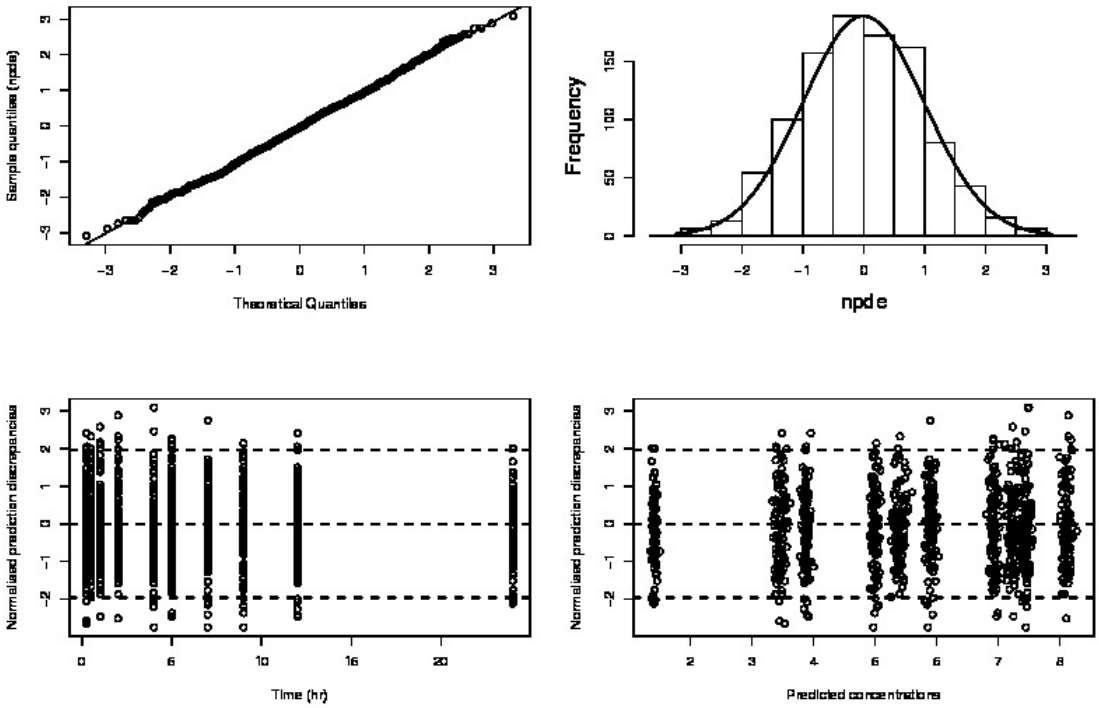

FIGURE 2. *Plots provided by the* npde *library for* V_{true}. *Top: QQ-plot of the distribution of the* npde *versus the theoretical* $\mathcal{N}(0,1)$ *distribution (left). Histogram of the distribution of the* npde*, with the density of the standard Gaussian distribution overlaid (right). Bottom: scatterplot of* npde *as a function of time (left) and predicted concentrations (right); dashed lines show the lines y=0 and y=*$\pm$*1.96.*

 ISSN: 2102-6238

Figure 3 shows a plot of npde versus the decorrelated predictions for V_{true}. The decorrelation tends to spread out the data over the whole range of (decorrelated) predictions.

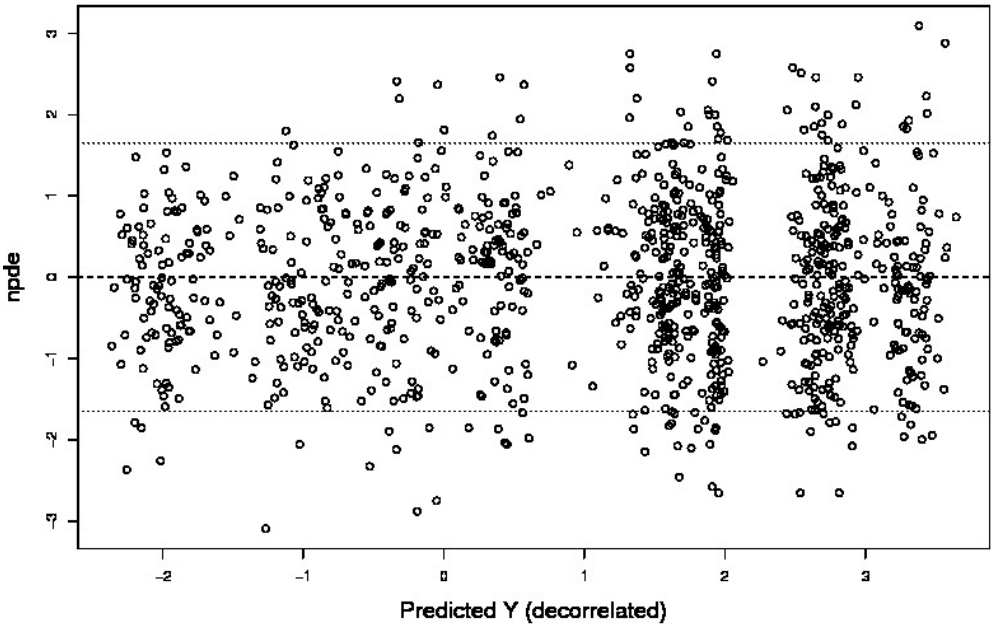

FIGURE 3. *Scatterplot of* npde *versus decorrelated predictions for* V_{true}. *Dashed lines show the lines* $y=0$ *and* $y=\pm 1.96$.

Figure 4 shows several representations of VPC. In the top panel, the left figure represents the observed data as points, dashed lines join the 2.5, 50 and 97.5th percentiles of the simulated data with a thicker line for the median. The right figure shows the same plot, but with a coloured area to make it easier to visualise where the bulk of the data is expected to lie. The bottom panel shows another representation of VPC: prediction bands (see methods) are plotted around the 2.5, 50 and 97.5th percentiles of the simulated data. The lines now represent the 2.5, 50 and 97.5th percentiles for the observed data instead of the simulated data (which we could also plot, but here lead to a rather busy plot and were omitted for clarity). The lower two figures differ only by overlaying the data or not. Different colours can be used for the extreme percentiles (2.5 and 97.5) and the PI around the median, both when representing the prediction intervals and when plotting the lines themselves. VPC with prediction bands as shown in the two lower plots are a very vivid representation of the data that can make it immediately clear in which areas the model still needs improvement. The four plots in figure 4 show different ways of comparing the model predictions to observed data, with additional information in the two lower ones. The two upper figures are identical save for the shaded area. Both have been used in the literature, but the right hand figure is more visually appealing and is increasingly replacing the former. The two lower figures include additional information, as the shaded areas can be used to compare observed percentiles to model predictions. Depending on the study design, the amount of data and their spread may clutter the

 ISSN: 2102-6238

graph, making the lower right-hand plot more appealing. Alternatively, only the observations outside the shaded areas may be plotted instead.

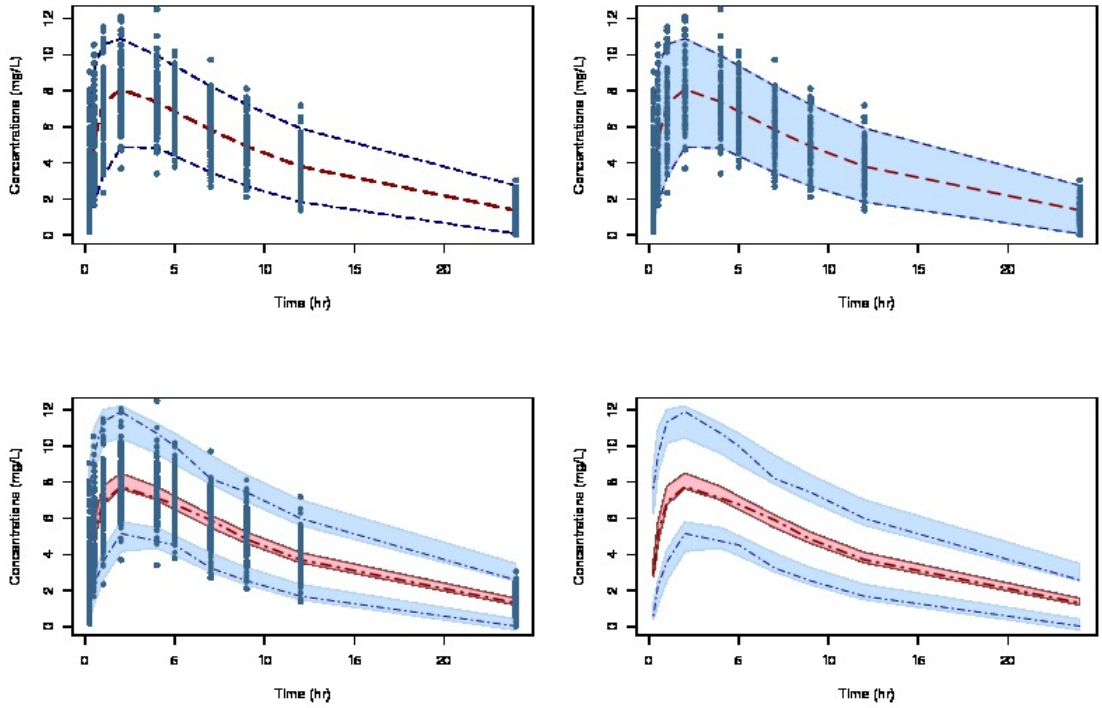

FIGURE 4. *VPC plots for V_{true}, with several representations. Top: 2.5 and 97.5$^{\text{th}}$ percentiles of the simulated data, shown as dashed lines (left) or coloured area (right); for both graphs the 50$^{\text{th}}$ percentile is shown as a thick dashed line and the observations are overlaid as dots. Bottom: 95% prediction intervals around 2.5, 50 and 97.5$^{\text{th}}$ percentiles of the simulated data shown as coloured areas; the 2.5, 50 and 97.5$^{\text{th}}$ percentiles of the observed data are plotted as dotted/dashed lines (– . –) with a thicker line for the median.*

 ISSN: 2102-6238

Prediction bands can also be obtained to complete the scatterplots of pd and npde versus time or predictions. As for VPC, we compute pd and npde for each simulated dataset to obtain prediction bands around selected percentiles of the observed pd and npde; the results are shown in figure 5, for pd (top) and npde (bottom), with similar conventions as for figure 4. Again, adding the area in which the median and limits of the 95% interval are expected to be found is very visually appealing. Of the two sets of plots, when model misspecification is being investigated, pd can be preferred over npde. Indeed, npde have been transformed to remove the correlation between the samples taken in the same individual, while pd remain closer to the original data.

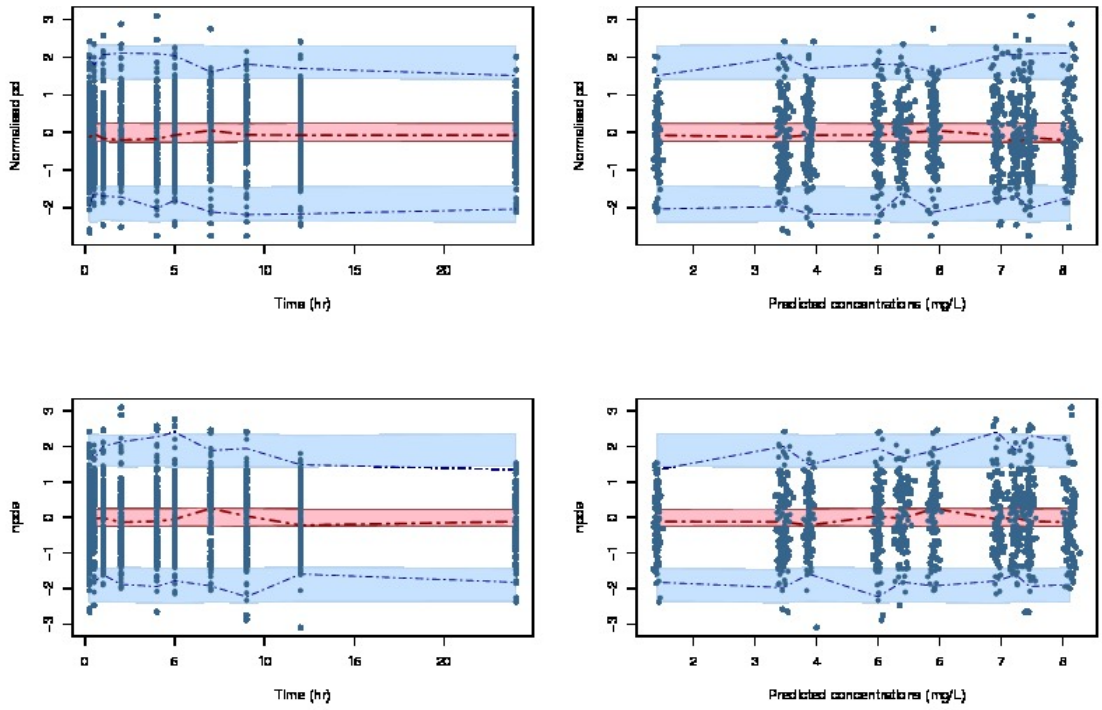

FIGURE 5. *Graphs of* pd *(top) and* npde *(bottom) for* V_{true}*, plotted versus time (left) and predicted concentrations (right), with prediction bands (see legend of Figure 4 and text for description).*

 ISSN: 2102-6238

Figure 6 shows the VPC with prediction bands for the four datasets. We chose not to show the individual observations to reduce the clutter, but the same plots could include individual datapoints. The VPC are able to highlight all the model misspecification we simulated.

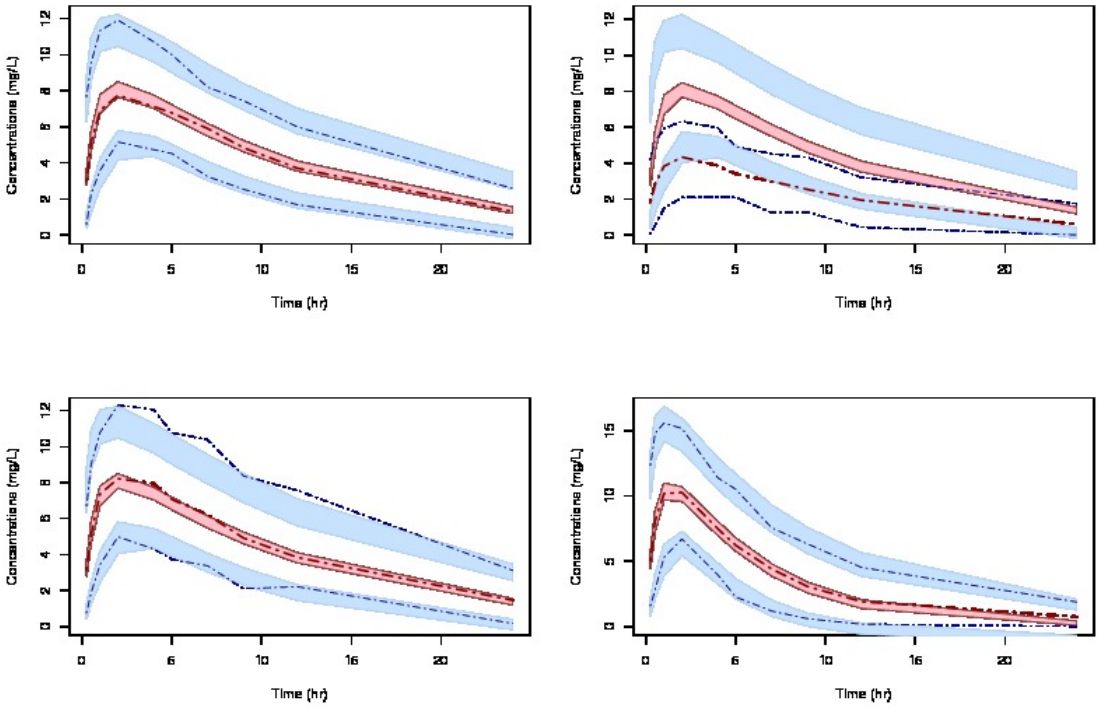

FIGURE 6. *VPC with prediction bands, for datasets* V_{true} *(upper left),* V_{bioavail} *(upper right),* V_{IIV} *(lower left),* V_{2cpt} *(lower right).*

 ISSN: 2102-6238

Figure 7 shows the pd versus time with prediction bands for the four datasets (again without observations overlaid). Model misspecification is again very clear for V_{bioavail} because of the large change in the volume of distribution. For V_{IIV}, there are too many extreme values of pd but no trend in time, suggesting a problem with the variability model which was maybe not as readily apparent with VPC; spikes outside the prediction bands can be seen around the same time as the spikes in the VPC plot. Finally, for $V_{2\text{cpt}}$, structural model misspecification appears as a trend in time only visible at the latest time.

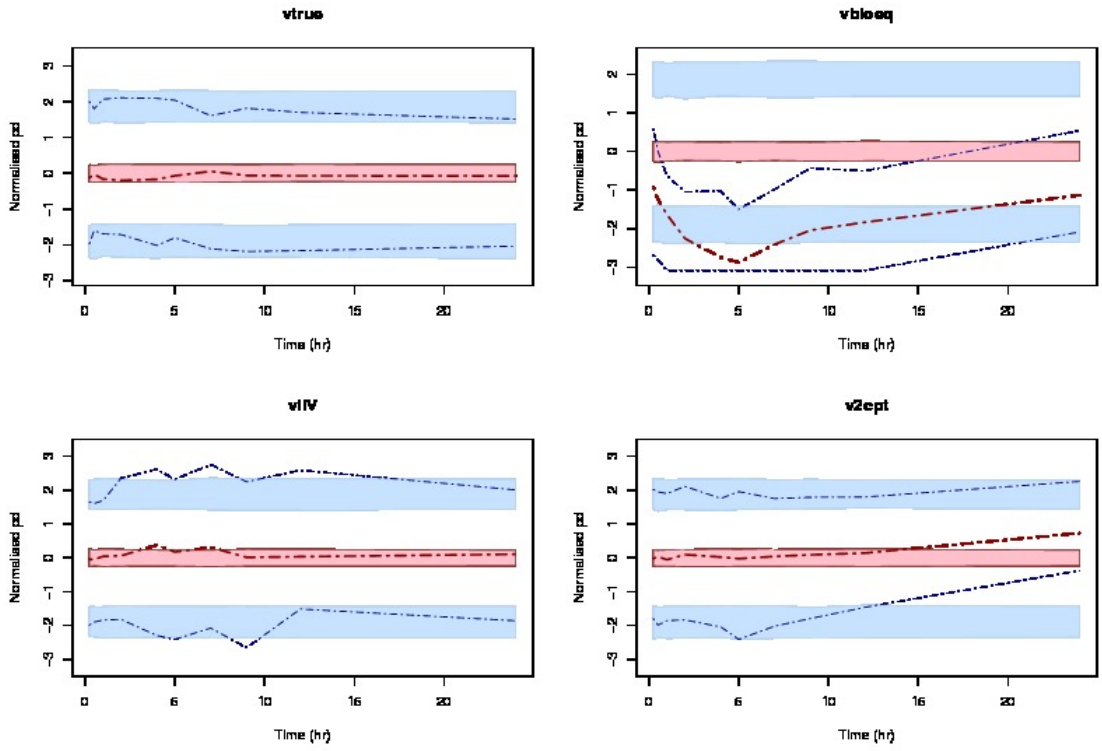

FIGURE 7. *Plot of* pd *versus time with prediction bands, for datasets* V_{true} *(upper left),* V_{bioavail} *(upper right),* V_{IIV} *(lower left),* $V_{2\text{cpt}}$ *(lower right).*

4. Discussion

A number of metrics have been described in [7]. We briefly recalled the most promising of them and illustrated them. The purpose of the present paper is not to evaluate or compare the different metrics, but rather to present examples of their use for model evaluation, and to show how they can be made particularly striking by adding prediction bands around important features of the model. A single dataset was simulated for each simulation settings. Therefore, this presentation is by no means exhaustive nor does it claim to provide a definitive methodology for model evaluation. Some metrics, such as the weighted residuals (called WRES in NONMEM), also

 ISSN: 2102-6238

called standardised prediction errors [40], were not included on purpose for the present illustration. Although they have been extensively used, their ineffectiveness has indeed already been amply demonstrated [33, 8]. WRES rely on a first-order approximation of the model, as do another metric, the conditional weighted residuals CWRES [22]. npde are residuals based on the whole predictive distribution for each observation, and avoid the linearisation problem. Recently, Laffont and Concordet proposed a metric termed GUD for Global Uniform Distance [27], as well as a circular diagnostic graph. This metric requires extensive simulations to define a prediction region not unlike the prediction bands around the boundaries that have been proposed for VPC and applied also to pd and npde in the present paper.

In fact, with the improvement in computer power, several metrics using extensive simulations to approximate the distribution of a statistic under the model being tested have been made available. Normalised prediction distribution errors, npde, are an example of a larger class of statistics called posterior predictive check (PPC). They were first proposed for non-parametric estimation methods by Mesnil et al. [34]; in that setting, they can be computed exactly because the distribution of random effects is discrete. They were more recently extended to parametric estimation methods where they are computed using Monte-Carlo methods [33], and improved by taking correlations into account [7]. The idea can be traced back to the Bayesian concept of using the whole predictive distribution for model evaluation, as discussed in Gelfand et al [15] and Gelman et al [16]. In the field of pharmacokinetics/pharmacodynamics, Girard et al. used the predictive distribution of compliance pattern to compare models [18]; more recently, Yano et al. formalised the concept of PPC and evaluated a number of standard statistics [43]. PPC extend readily to non-continuous data, and a nice example of VPC applied to categorical data can be found for instance in [19], where the evolution with time of the probability of experiencing hand and foot syndrome after exposition to two anticancer agents was modelled using a proportional odds model and evaluated with odd-type VPC. An application of npde to count data has also demonstrated the efficiency of this metric even with non-continuous data [38].

Posterior predictive checks come naturally in the Bayesian framework, which uses the posterior distribution resulting from Bayesian inference to compute the distribution of a future observable quantity, conditionally on observed values [37]. Posterior predictive p-values measuring the discrepancy between the data and a model have been proposed which reconcile the Bayesian analysis and the frequentist notion of p-values [1]. Different types of p-values have been proposed, based on different distributions, such as prior predictive distribution [6], posterior predictive distribution [37, 32], partial predictive distribution or conditional predictive distribution[1]. The properties of the different p-values were studied in several examples, and Robins et al in a companion paper studied their asymptotic distributions[36]. Their results suggest a superiority of conditional predictive p-values and partial posterior predictive p-values, but admit that the computation of these p-values may be difficult; their extension to nonlinear mixed-effect models would be a subject worthy of investigation but is far beyond the scope of the present paper. Concerning posterior predictive p-values, Bayarri and Berger [1] argue that the *plug-in* approach, which consists in using point estimates and disregarding uncertainty, avoids a double-use of the data occurring when using the full posterior distribution, but point out the resulting p-value may be conservative. This finding was corroborated by Yano et al. who in their PK/PD simple example found the *plug-in* approach to be appropriate, although they caution that this could depend on the design of the study and the magnitude of interindividual variability [43]. A more in-depth

 ISSN: 2102-6238

discussion on this topic can be found in [33].

Simulation-based metrics require to be able to simulate all the features of the data. As noted by Karlsson and Savic [25], this may pose several problems. First, some observational designs are not readily amenable to simulations. Second, clinical trials often include drop-outs, censored or missing data, and the construction of diagnostic graphs need to take these processes into account in the model. In particular, all these simulation-based metrics implicitely assume the data does not contain censored data, such as data under the limit of quantification (BQL data); if the amount of BQL data is large at certain time points, it can be worthwhile to compare the percentage of BQL data predicted by the model to the percentage observed in the dataset [8]. Drop-out processes can be modelled to correct for model misspecification [24], but patterns will appear in plots of the npde or in the VPC when the data is not missing completely at random (MCAR). When data is missing at random (MAR), imputation of unobserved values may be used to complete diagnostic plots and remove these trends, as well as provide insight into model features [17] and this has been shown to work well with npde and VPC [28].

Models should be evaluated by different tools; in the quick illustration presented here, we could see how VPC and/or normalised pd with prediction intervals can pinpoint model deficiencies, while npde or PI-NPC provide an overall test which can be more sensitive than the human eye. Indeed, in the present illustration, npde were able to correctly detect all the model misspecifications we simulated in our examples, when using the global test with Bonferroni correction, and examining the individual tests provided some insight as to which feature of the model should be improved. Plots of pd versus time with prediction bands and VPC plots were also very sensitive to the various model misspecifications simulated here. With the Kolmogorov-Smirnov test on npde, on the other hand, we did not detect a problem with the dataset where the IIV was misspecified. However we need to recall that we only changed the IIV from 13% to 25% and only on one parameter, which is a very small change that may be both difficult to detect in practice and unlikely to be clinically significant. PI-NPC or their decorrelated version appeared on the other hand sensitive to misspecification in IIV, but here did not indicate a problem with V_{2cpt} where the structural model was misspecified. These differences were somewhat surprising since there is a close relationship between VPC, where we consider the position of the observation within the distribution of the simulated concentrations, and pd which are the quantiles of the observations in this distribution. This result may therefore be related to the different statistical tests involved, as we have already seen differences between using a Kolmogorov-Smirnov test or a global test involving three sub-tests. Examining the strengths and weaknesses of the different tests would require a more in-depth investigation than the simple illustration performed in the present review. In addition, in real datasets, model misspecifications may occurr simultaneously for several model features, which may make them more difficult to identify; in addition, the presence of outliers or large amounts of data may render the tests (especially the normality test) inordinately sensitive in practice, so that the results of the statistical tests should probably then be used more as an indication to guide model building.

VPC and PI-NPC require rather homogenous designs for all subjects, while pd and npde take naturally into account individual differences in designs. When sampling times differ, binning can be an option. When doses differ, or when covariates enter the model, one solution for VPC is to stratify, but the amount of information per stratum will then decrease, and the VPC may become less informative. To solve this conundrum, it has recently been proposed to normalise VPC using

 ISSN: 2102-6238

the median predicted value at each time point (or within in binning interval), thus producing Prediction Corrected VPC (PC-VPC) [3]. PC-VPC partly correct for design heterogeneity, and are more efficient to detect model misspecification; we could also propose a PI-NPC test for the prediction corrected VPC. However, contrary to pd and npde there does not appear to be a theoretical rationale for the correction proposed; PC-VPC also do not take into account within subject correlations. In the present paper, the designs we simulated were homogenous, with the same dose and sampling times for all individuals and no covariate in the model, therefore VPC and PC-VPC are the same, since the population prediction for each time-point is the same across all subjects and the ratio involved in the computation of PC-VPC is equal to 1 for all observations. In particular, we chose to simulate the same dose for all subjects, to avoid having to split the graphs for VPC. It would therefore be interesting to evaluate PC-VPC as well as the other diagnostic tools in a more challenging setting.

Plots of residuals are traditionally plotted versus time and predicted concentrations. With weighted residuals or npde however, we could consider decorrelating the variables plotted on the X-axis. For predicted concentrations, the variance-covariance matrix of the individual vector of predictions can be used to perform the decorrelation. We show an example of such a graph in figure 3 for V_{true}, and it would be interesting to study in more detail how these graphs could be used for model diagnostics. In the present work, they performed similarly to the graphs using non-decorrelated predictions: a clear downward trend appeared for $V_{bioavail}$, but the other two model misspecifications were not readily apparent (data not shown). Diagnostic plots are not limited to the plots versus time and predictions. Commonly used plots include graphs of the predictions versus the observations, individual fits, as well as plots of absolute prediction errors. In addition, other elements to help diagnose models are the standard errors of estimation (SE), which are reported by the software and can sometimes help to detect overparameterisation. The modeller should therefore select features of the model which are important for the purpose of the analysis, and choose diagnostic tools to evaluate these aspects. For instance, the global test proposed to test the distribution of the npde is a combination of three tests; the Wilcoxon test assessing whether the mean of the npde is significantly different from 0 can be used to evaluate whether the structural model is appropriate, and can be completed by an absence of trend in the graph of npde versus time to check whether this hypothesis holds with time. This can also be seen through a VPC with the simulated median superimposed. In addition, the validity of certain diagnostics may depend on design, as there is evidence that the usefulness in particular of individual predictions or individual residuals may be decreased in the presence of significant shrinkage [25].

Outliers can be seen in most diagnostic graphs, especially with real data. Although the ability of the different approaches to detect outliers or influential data has not to our knowledge been formally studied, Semmar et al. have shown that npde values of large magnitude were correlated with outlier data identified by *a priori* multivariate techniques [39]. VPC and plots of the npde or pd versus time can therefore be used to gain an impression of outliers in a dataset. However, this approach probably makes more sense when applied to external evaluation, since in internal evaluation the outlier value is included in the dataset used to estimate the parameters.

A number of tools providing diagnostic plots are available to modellers. For users of the Monolix software, a large number of diagnostic plots are output for each run, including VPC plots as well as histograms and QQ-plots of the distribution of npde. For users of the NONMEM

 ISSN: 2102-6238

software, libraries for R such as Xpose [23], WfN [20] and npde [10] are very useful, and the latest version of NONMEM (version 7) outputs VPC and npde. Xpose can be combined with PsN (Perl speaks Nonmem) [29] to produce VPC plots as well as coverage plots for different intervals. Prediction bands for VPC or other metrics can be produced with R code, as we did in the present examples. Computing the prediction bands for the pd and especially the npde can be somewhat computationally cumbersome, since it requires to compute the metrics for each of the simulated datasets, but if this is really a challenge then the prediction bands can be computed in a subset, for instance 200, of the simulated datasets.

Internal evaluation applies the metrics described above to the dataset used to build the model, while external evaluation uses data not involved in parameter estimation. The metrics illustrated here can be applied indifferently for internal or external evaluation. Used for internal evaluation, they provide model diagnostics to guide model building and assessment. External evaluation is considered to be the most stringent, allowing to evaluate the predictive ability of the model in an independent dataset. However, V should be relatively similar to B so that the evaluation remains meaningful. If the number of subjects is large enough, data-splitting approaches can be used to divide the full dataset in a building and a validation subset. Cross-validation using the metrics described above, by repeatedly splitting the data in building and evaluation datasets, can also be used to assess the true predictive capacity of the model on a separate sample. We should of course always keep in mind that the approaches reviewed here only allow to reject a model and never to accept it. An alternative would be to develop tests of model adequacy, using for instance bioequivalence principles.

Acknowledgement

The authors would like to thank the two anonymous referees for their constructive comments which helped put this work in a larger perspective, and are indebted to one of the referees for suggesting to use decorrelated predictions instead of the standard predictions in the residual graphs (figure 3).

References

[1] M Bayarri and J Berger. P values for composite null models. *Journal of the American Statistical Association*, 95:1127–42, 2000.

[2] TR Belin and DB Rubin. The analysis of repeated-measures data on schizophrenic reaction times using mixture models. *Statistics in Medicine*, 14:747–68, 1995.

[3] M Bergstrand, AC Hooker, JE Wallin, and MO Karlsson. Prediction Corrected Visual Predictive Checks. *American Conference on Pharmacometrics, October 4-7, 2009, Mashantucket, USA*, 2009.

[4] A Boeckmann, L Sheiner, and S Beal. *NONMEM Version 5.1*. University of California, NONMEM Project Group, San Francisco, 1998.

[5] G Box. Science and statistics. *Journal of the American Statistical Association*, 71:791–9, 1976.

[6] GEP Box. Sampling and Bayes inference in scientific modeling and robustness. *Journal of the Royal Statistical Society A*, 143:383–430, 1980.

[7] Karl Brendel, Emmanuelle Comets, Céline Laffont, Christian Laveille, and France Mentré. Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. *Pharmaceutical Research*, 23:2036–49, 2006.

 ISSN: 2102-6238

[8] Karl Brendel, Emmanuelle Comets, Céline Laffont, and France Mentré. Evaluation of different tests based on observations for external model evaluation of population analyses. *Journal of Pharmacokinetics and Pharmacodynamics*, 37:49–65, 2010.

[9] Karl Brendel, Céline Dartois, Emmanuelle Comets, Annabelle Lemenuel-Diot, Christian Laveille, Brigitte Tranchand, Pascal Girard, Céline Laffont, and France Mentré. Are population pharmacokinetic and/or pharmacodynamic models adequately evaluated? A survey of the literature from 2002 to 2004. *Clinical Pharmacokinetics*, 46(3):221–234, 2007.

[10] Emmanuelle Comets, Karl Brendel, and France Mentré. Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: the npde add-on package for R. *Computer Methods and Programs in Biomedicine*, 90:154–66, 2008.

[11] Committee for Medicinal Products for Human Use, European Medicines Agency. *Draft guideline on reporting the results of population pharmacokinetic analyses*. EMEA, 2006.

[12] M Davidian and D Giltinan. *Nonlinear models for repeated measurement data*. Chapman & Hall, London, 1995.

[13] E Ette and PJ Williams. *Pharmacometrics: the science of quantitative pharmacology*. Wiley-Interscience, Hoboken, New Jersey, 2007.

[14] Food and Drug Administration. *Guidance for Industry: Population Pharmacokinetics*. FDA, Rockville, Maryland, USA, 1999.

[15] AE Gelfand, DK Det, and H Chang. *Bayesian statistics*. Oxford University Press, Oxford, 1992.

[16] A Gelman, J Carlin, H. Stern, and D Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[17] A Gelman, I Van Mechelen, G Verbeke, D Heitjan, and M Meulders. Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics*, 61:74–85, 2005.

[18] P Girard, T Blaschke, H Kastrissiosr, and L Sheiner. A Markov mixed effect regression model for drug compliance. *Statistics in Medicine*, 17:2313–33, 1998.

[19] E Hénin, B You, E VanCutsem, P M Hoff, J Cassidy, C Twelves, K P Zuideveld, F Sirzen, C Dartois, G Freyer, M Tod, and P Girard. A dynamic model of hand-and-foot syndrome in patients receiving capecitabine. *Clinical Pharmacology & Therapeutics*, 85:418–25, 2009.

[20] N Holford. *Wings for Nonmem, http://wfn.sourceforge.net*. University of Auckland, Auckland, New Zealand, 2000.

[21] N Holford. The Visual Predictive Check: superiority to standard diagnostic (Rorschach) plots. *14$^{\text{th}}$ meeting of the Population Approach Group in Europe, Pamplona, Spain*, page Abstr 738, 2005.

[22] A Hooker, C Staatz, and MO Karlsson. Conditional Weighted Residuals (CWRES): a model diagnostic for the FOCE method. *Pharmaceutical Research*, 24:2187–97, 2007.

[23] E Jonsson and M Karlsson. Xpose–an S-PLUS based population pharmacokinetic/pharmacodynamic model building aid for NONMEM. *Computer Methods and Programs in Biomedicine*, 58(1):51–64, 1999.

[24] M Karlsson and N Holford. A tutorial on visual predictive checks. *17$^{\text{th}}$ meeting of the Population Approach Group in Europe, Marseille, France*, page Abstr 1434, 2008.

[25] MO Karlsson and RM Savic. Diagnosing model diagnostics. *Clinical Pharmacology & Therapeutics*, 82:17–20, 2007.

[26] E Kuhn and M Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, 49:1020–1038, 2005.

[27] C Laffont and D Concordet. A new exact test to globally assess a population pk and/or pd model. *18$^{\text{th}}$ meeting of the Population Approach Group in Europe, St-Petersburg, Russia*, page Abstr 1633, 2009.

[28] A Lemenuel-Diot, C M Laffont, R Jochemsen, and E Foos-Gilbert. Evaluation of model of heart rate during exercise tolerance test with missing at random dropouts. *16$^{\text{th}}$ meeting of the Population Approach Group in Europe, Copenhague, Denmark*, page Abstr 1227, 2007.

[29] L Lindbom, P Pihlgren, and EN Jonsson. PsN-Toolkit–a collection of computer intensive statistical methods for non-linear mixed effect modeling using NONMEM. *Computer Methods and Programs in Biomedicine*, 79:241–57, 2005.

[30] M Lindstrom and D Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46:673–87, 1990.

[31] A Mallet. A maximum likelihood estimation method for random coefficient regression models. *Biometrika*, 73:645–56, 1986.

 ISSN: 2102-6238

[32] XL Meng. Posterior predictive p-values. *Annals of Statistics*, 22:1142–60, 1994.

[33] F Mentré and S Escolano. Prediction discrepancies for the evaluation of nonlinear mixed-effects models. *Journal of Pharmacokinetics and Biopharmaceutics*, 33:345–67, 2006.

[34] F Mesnil, F Mentré, C Dubruc, JP Thénot, and A Mallet. Population pharmacokinetics analysis of mizolastine and validation from sparse data on patients using the nonparametric maximum likelihood method. *Journal of Pharmacokinetics and Biopharmaceutics*, 26:133–61, 1998.

[35] J Pinheiro and D Bates. Approximations to the log-likelihood function in the non-linear mixed-effect models. *Journal of Computational and Graphical Statistics*, 4:12–35, 1995.

[36] J Robins, A van der Vaart, and V Ventura. Asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Society*, 95:1143–56, 2000.

[37] DB Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12:1151–72, 1984.

[38] RM Savic and M Lavielle. Performance in population models for count data, part ii: A new saem algorithm. *Journal of Pharmacokinetics and Pharmacodynamics*, 36:367–79, 2009.

[39] N Semmar, S Urien, B Bruguerolle, and N Simon. Independent-model diagnostics for a priori identification and interpretation of outliers from a full pharmacokinetic database: correspondence analysis, mahalanobis distance and andrews curves. *Journal of Pharmacokinetics and Pharmacodynamics*, 35:159–83, 2008.

[40] S Vozeh, P Maitre, and D Stanski. Evaluation of population (NONMEM) pharmacokinetic parameter estimates. *Journal of Pharmacokinetics and Biopharmaceutics*, 18:161–73, 1990.

[41] J Wilkins, M Karlsson, and N Jonsson. Patterns and power for the visual predictive check. *15*$^{\text{th}}$ *meeting of the Population Approach Group in Europe, Brugges, Belgium*, page Abstr 1029, 2006.

[42] R Wolfinger. Laplace's approximation for nonlinear mixed models. *Biometrika*, 80:791–5, 1993.

[43] Y Yano, S Beal, and LB Sheiner. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. *Journal of Pharmacokinetics and Pharmacodynamics*, 28(2):171–192, Apr 2001.

 ISSN: 2102-6238