




## Review Article

Theme: Celebrating Women in the Pharmaceutical Sciences

Guest Editors: Diane Burgess, Marilyn Morris and Meena Subramanyam

# Developing Tools to Evaluate Non-linear Mixed Effect Models: 20 Years on the npde Adventure

Emmanuelle Comets<sup>1,2,3</sup> and France Mentré<sup>1</sup> 

Received 4 December 2020; accepted 16 April 2021; published online 19 May 2021

**Abstract.** This article revisits 20 years of our work in developing evaluation tools adapted to non-linear mixed effect models. These hierarchical models involve a large number of assumptions concerning the structural evolution of the outcomes, the link between different outcomes, the variabilities in the parameters and model evaluation aims at assessing these various components, both to help guide the model building and to communicate on model adequacy for a given purpose. During our career, we have developed and extended simulation-based evaluation tools called normalised prediction discrepancies (npd) and normalised prediction distribution errors (npde), providing informative diagnostics through graphs and tests.

**KEY WORDS:** mixed effect models; model diagnostics; model evaluation; NLMEM; npde.

## INTRODUCTION

The last 4 decades have seen an exponential increase in the use of non-linear mixed effect models (NLMEM) in pharmacology. They were first developed to characterise the pharmacokinetics (PK) and pharmacodynamics (PD) of new medications and extensively used in drug development. Since then their usage has expanded, to help design clinical trials by optimising the number and arrangement of samples to maximise information, to therapeutic drug monitoring by individualising drug regimens, to managing long-term therapy by incorporating disease models. Methodological developments kept pace as the models progressively became more complex, describing multiple simultaneous outcomes, diversifying response types to describe continuous, time to event (TTE) or discrete data, incorporating different levels of variability to account for variations both within and between subjects or clusters, modelling underlying disease evolution, etc. While the first algorithms involved approximations to the log-likelihood to allow the estimation of population parameters (1), improvements in computing power have made it possible to run the new sophisticated statistical methods developed both in frequentist analyses and Bayesian approaches (2). Specialised software dedicated to NLMEM now

include the pioneer NONMEM (3) and Monolix, the first to implement the stochastic expectation-maximisation (EM) algorithm (4). More general software with NLMEM support include the Bayesian Stan (5), and packages for mainstream statistical software like R and Julia are also available, such as nlme (6), lme4 (7), saemix (8), nlmixr (9) and Pumas (<https://pumas.ai>).

In parallel to the question of “how to estimate the parameters?”, numerous methodological works also tackled questions such as “what data are needed?”, which sprung a whole area of research into optimal design (10); “what to do with the results?”, with applications such as therapeutic drug monitoring which formed the initial impetus for developing population methods in the first place (11); and “what confidence do we have in them?”. This last question of model evaluation was also addressed very early by Lewis Sheiner and Stuart Beal in 1981 (12) and is crucial to the use of NLMEM. In fact, model evaluation was integrated into the guidelines on population pharmacokinetic analyses, first in the European guidance (13) and more recently into the revised version of the guideline by the Food and Drug Administration (14).

In this article, we will focus on our contribution to this field, the normalised prediction discrepancies (npd) and normalised prediction distribution errors (npde), and how they have become part of the gold standards in the evaluation of these models. We will first recall the history of model evaluation and how the npde have taken their place amongst the other evaluation tools. We will present the concepts of normalised prediction discrepancies (npd) and normalised prediction distribution errors (npde) in the “Concepts”

Guest Editors: Diane Burgess, Marilyn Morris and Meena Subramanyam

<sup>1</sup> Université de Paris, INSERM, IAME, F-75006, Paris, France.

<sup>2</sup> INSERM, CIC1414, Université Rennes-1, 35000, Rennes, France.

<sup>3</sup> To whom correspondence should be addressed. (e-mail: emmanuelle.comets@inserm.fr)

section of the manuscript and the npde package, where we implement our tools and extensions, in “The npde Package” section. We will finish in “The npd Following” section by looking at the articles, both methodological and applied, citing 5 foundation papers on npde we wrote along this adventure.

## HISTORY

### Evaluating Non-linear Mixed Effect Models

The issue of model evaluation was first raised by the pioneers of pharmacometrics in an article touching on various problems including evaluation metrics and experimental conditions (12). A natural criterion to assess the performance of an analysis is to measure the prediction error, representing the difference between the true value and the estimate, and to consider both its systematic component, which may indicate bias or systematic model misspecification, and its magnitude, which is a measure of imprecision. In longitudinal analyses, these prediction errors may vary with time or covariates, and graphical diagnostics are used to examine trends and suggest model improvements. However, it was recognised early on in the development of population analyses that classical residuals defined simply as the difference between predictions and observations suffered from several shortcomings in NLMEM. A first issue lies in the definition of the prediction as population or individual predictions, the duality reflecting the hierarchical nature of NLMEM. Population residuals were initially defined using the prediction for the ‘typical’ population parameters, and assess the model as a whole; their magnitude results from both between and within subject variability. Individual residuals require the estimation of individual parameters through Bayesian approaches, are more focused on the residual error model, and have been shown to be very sensitive to shrinkage induced by low information content in the design for some or all subjects (15). Shrinkage describes the regression of the individual parameters towards the population estimates during the estimation and increases with sparse or poorly informative sampling. A second issue is potential heteroscedasticity in the residual error model, manifesting itself by an expected magnitude of the residuals varying across the range of observations, and which may be corrected by weighting the residuals using the expected variance of the prediction, leading to weighted residuals (WRES) available for instance in NONMEM output and individual weighted residuals (IWRES) used to assess the residual error model. Because of the bias introduced by the first-order approximation and the fact that their true distribution is unknown, the performance of WRES for model evaluation in NLMEM has been consistently shown to be poor (16,17).

We were the first to recognise the importance of explicitly accounting for the non-linearity in the structural model, by considering the entire predictive distribution of an observation. In attempting to evaluate a population PK analysis of mizolastine in sparse data using the non-parametric maximum likelihood approach (18), we realised that the discrete nature of non-parametric distributions allowed an explicit computation of the quantile of an observation in the predictive distribution, a quantity we

termed prediction discrepancy (pd). We showed that by construction, these quantiles distributed evenly within a cumulative distribution and thus followed an overall uniform distribution provided that the model used to produce the predictive distributions is correct. We also proposed to transform them to a normal distribution, using the normal inverse cumulative density function, into normalised prediction discrepancies (npd) which could then be used to produce evaluation graphs similar to those obtained with standard weighted residuals. With a more usual parametric assumption for the random effects, the computation involved in definition of the npd no longer has an explicit solution but can be solved numerically by Monte-Carlo simulations (16), and we showed that the npd was much more reliable than WRES. Prediction discrepancies can be considered as an extension to NLMEM of the quantile residuals proposed by (19) as an alternative to Pearson or deviance residuals in generalised logistic regression. An issue that remained was that in NLMEM, repeated observations collected for a given individual are correlated through this individual’s parameters. This leads to an inflation in the type I error when comparing the npd to their theoretical distribution (16). The next step was therefore to account for this correlation, which led to the development of normalised prediction distribution errors, npde, which can be compared to their theoretical distribution through a combined test (17). This approach has been shown to perform well in simulation studies, with an adequate control of the type I error (20).

In these seminal articles, we showed that npd and npde are a form of posterior predictive check (PPC), an approach commonly used in Bayesian approaches and applied to the evaluation of population PK/PD analyses by Yano *et al.* (21). The idea of the PPC consists in choosing a statistic that can be computed from the observed data and comparing it to the distribution of the statistic generated under a given model to check whether it is compatible. With npd, we derive residuals as quantiles of what can be considered as a PPC of the observations. Another related approach is the visual predictive check (VPC), where the distribution of the observations is generated for the population using PPC and directly compared to observed percentiles of the data (22). Model evaluation is primarily communicated through graphical analysis, and the VPC are often used as an evaluation tool in pharmacometric analyses, as quantifying model misspecifications through associated numerical predictive checks (NPC) requires pre-specifying intervals of interest. One problem with VPC is that when the design is different across subjects, for instance, if they receive different doses, the distribution of observations over the whole set of subjects represents a mixture of potentially very different populations, whereas npde naturally accounts for differences between designs and covariates through the predictive distributions associated to a given observation. Prediction-corrected VPC (pcVPC) (23), which bin the observations before weighing them according to the expectation of the predictions in the bin, partially corrects for this but still imply approximations which may not be fully adequate for very non-linear models.

In parallel, a number of papers suggested improvements to WRES, most notably through better approximations to the prediction. Using Monte-Carlo simulation to derive population weighted residuals (PWRES) helped alleviate the bias

with WRES, and using a first-order conditional approximation (FOCE) to adjust WRES to the individual predictions of mean and variance led to residuals called conditional weighted residuals (CWRES) which provide similar evaluation graphs as npde, although their distribution is strictly known only if the structural model is linear (24,25).

### Community Uptake

The new model evaluation tools were soon implemented in software. We implemented the computation of npd and npde in a library called npde (26) for the R statistical software (27) in 2007. Monolix (4) implemented npde in version 2.3 (November 2007), along with VPC and residual plots. Monolix produces similar graphs as the npde package, as does the sister package saemix which performs estimation of parameters in NLMEM in R via the SAEM algorithm (8). The first version of NONMEM (3) to include npde in the output was version 7.1.0 in 2009, and NONMEM users can use the Xpose package to produce evaluation plots for a variety of metrics including npde, individual and population residuals, CWRES or VPC (28). The Pumas software, using the programming language Julia, also includes npde as an evaluation tool (<https://pumas.ai/>).

Diagnostic graphs benefitted from their more widespread availability in software, and they are now recognised as an integral part of the model building process. To help make sense of the proliferation of different evaluation tools, a Model Evaluation working group, led by Pr Mentré, was established by the International Society of Pharmacometrics (ISoP) Best Practice Committee (<http://go-isop.org/committees/standards-best-practices-committee/>), to create guidance concerning their definition, scope and usage. This group produced a first tutorial in 2017 illustrating the graphical evaluation process for NLMEM describing continuous responses (29), examining the merits and limits of evaluation tools using two simulated examples. Prediction-based residuals such as PWRES and CWRES could be used to assess the general shape of the structural model, and IWRES were useful to check the residual error model. However, graphs using PWRES or IWRES were not effective to detect misspecifications in the variability structure. Individual estimates of the random effects can be used to orient covariate selection and suggest an appropriate correlation structure. However, evaluation graphs based on individual predictions lose informativeness in the presence of high shrinkage (29), with Karlsson and Savic suggesting a shrinkage of 50% (when defined using a ratio of variance) introduces or masks model misspecifications (15). Simulation-based residuals were generally good at detecting the different types of model misspecifications. In the presence of covariates or heterogeneous designs, pcVPC should be performed instead of standard VPC to take these features into account. The simulation-based metrics can also be stratified to examine the impact of covariates and orient covariate selection. Recognising their complementary nature, a core set of evaluation graphs was therefore recommended to systematically examine the different features of NLMEM. The evaluation stresses the robustness of simulation-based tools such as npd, npde and VPC, although they may require more computational effort to implement the CWRES being the

only prediction-based residual which could compare in terms of versatility to detect model misspecifications.

### Recent Extensions

In PK/PD applications, it is not uncommon to encounter data below the limit of quantification (LOQ), such as low concentrations of a drug, or undetectable viral loads. Common approaches to deal with these BLQ data in model evaluation included imputing them to a pre-determined value, removing them from the evaluation graphs, at the risk of skewing them, or reproducing the censoring process in the simulated data (30). In Monolix, BLQ data are imputed to the individual predictions under the model, which is a sensible approach but advantages the model used for the prediction. We proposed in (31) a different imputation approach to define prediction discrepancies for a censored observation, arguing that whereas it may be difficult to sample within the distribution of the observations, the distribution of the pd. under the model is known to be the uniform distribution under the null hypothesis that the model is correct. We therefore proposed to sample the pd. for a censored observation within a uniform distribution over an interval depending on the predicted probability of being BLQ for that observation, given the tested model. The predictive distribution over the censoring interval can then be used to impute the missing observation in the graphs.

The same approach was extended to right-censored or interval-censored TTE data, where the predictive distribution of the survival time is used to impute npd and event times for censored data (32).

Other types of data that can be modelled in NLMEM include count and ordinal data, and graphical evaluation is currently limited for these outcomes to overall VPC. Extensions to categorical data have been implemented in the evaluation graphs produced by Monolix (4) and are currently being formalised and implemented in the npde package.

## CONCEPTS

### Models and Notations

Following the definition set out in (17), let  $y_{i,j}$  the  $j$ th longitudinal observation recorded in subject  $i$  at time  $t_{i,j}$ . We define  $y_i$  as the vector of the  $n_i$  longitudinal measurements  $\{y_{i1}, y_{i2}, \dots, y_{in_i}\}$  for this subject  $i$ . When  $y$  is a continuous variable, we model the observation  $y_{i,j}$  through a known function  $f$  (possibly non-linear) of the structural model and a measurement error  $\epsilon$  as:

$$y_{i,j} = f(t_{i,j}, \psi_i, z_i) + g(t_{i,j}, \psi_i, z_i) \epsilon_{i,j} \quad (1)$$

where  $g$  represents the variance of the residual error and both  $f$  and  $g$  are assumed to depend on individual parameters  $\psi_i$  and subject-specific covariates  $z_i$ . The error term  $\epsilon_{i,j}$  is most often described using a standard normal distribution  $\mathcal{N}(0, 1)$ , and  $g$  can be constant ( $g = \sigma$ ) or depend on the model predictions with additional parameters.

In the non-linear mixed effect setting, we estimate the distribution of the individual parameters  $\mathcal{D}_\psi$  and not the  $\psi_i$  themselves. Usually, a parametric form is posited for  $\mathcal{D}_\psi$ , and the individual parameters are then expressed as a function  $h$  of fixed effects  $\mu$ , including typical parameter values in the population and covariate effects, and random effects  $\eta_i$ , specific for each subject:

$$\psi_i = h(\mu, z_i, \eta_i) \tag{2}$$

where the distribution of  $\eta_i$  is usually Gaussian with variance-covariance matrix  $\Omega$  ( $\eta_i \sim \mathcal{N}(0, \Omega)$ ). Additional levels of variability can be included in the model to represent intra-subject or intra-cluster variability. This parametric assumption on the distribution  $\mathcal{D}_\psi$  can also be relaxed, and Mallet (33) showed that in this case  $\mathcal{D}_\psi$  becomes a discrete distribution consisting of at most  $N$  vectors of parameters with associated frequencies.

With discrete data, we write the model directly in terms of probability. For instance, assuming  $y$  takes its values in a discrete set of  $K$  possible values  $\{c_1, c_2, \dots, c_K\}$ , we can use logistic functions for  $f$  to model the probability of  $y_{i,j}$  taking the value  $c_k$  (or less than  $c_k$  depending on the model chosen) as a function of individual parameters and covariates:

$$p(y_{i,j} = c_k | \psi_i, z_i) = f_k(t_{i,j}, \psi_i, z_i) \tag{3}$$

Similar equations can be derived for count or for time-to-event (TTE) data by modelling the hazard function (32).

In a frequentist setting, maximum likelihood approaches are used to estimate the population parameters, which for the continuous model defined by Eqs. (1) and (2) includes the fixed effects, the variance of the random effects and the parameters of the residual error model  $\theta = \{\mu, \Omega, \sigma\}$ . We denote  $\mathcal{M}$  the model defined by the structural, variance and distribution models, as well as a vector of population parameters  $\theta$ .

### Computing Prediction Discrepancies and Prediction Distribution Errors

Mentré and Escolano define prediction discrepancies (denoted pd) as the quantile of an observation in its marginal predictive distribution (16):

$$pd_{ij} = F_{ij}(y_{ij} | \theta) = \int_0^{y_{ij}} p_i(y | \theta) dy \tag{4}$$

where  $p_i(y | \theta)$  represents the conditional density of  $y$  in subject  $i$  with respect to the population parameters. In NLMEM, Eq. 1 can be used to express the density conditional to the individual parameters, but these are unknown variables in the model and need to be integrated out. Expression 4 becomes, denoting  $p(y | \psi_i, \theta)$  the conditional density of the data given the individual parameters, and  $p(\psi_i | \theta)$  the conditional density of the individual parameters given the population parameters, expression 4 becomes:

$$pd_{ij} = F_{ij}(y_{ij} | \theta) = \int_0^{y_{ij}} \int p(y | \psi_i, \theta) p(\psi_i | \theta) d\psi_i \tag{5}$$

In a non-parametric setting, the discrete nature of VI/J transforms the inner integral in Eq. (5) in a discrete sum which can be explicitly calculated (18). In the general parametric case, however, computing the prediction discrepancies requires computing the integral in Eq. (5) by sampling a large number of  $\psi_i$  in  $\mathcal{D}_\psi$ .

In Fig. 1, we show this process schematically for a continuous observation. On the left, we represented the cumulative distribution for an observation  $y_{ij}$  at time  $t_{i,j}$ , obtained by simulating a large number of values from  $\mathcal{M}$  and taking the cumulative distribution of the predictive distribution. The lines in blue represent how to compute the quantile  $pd_{ij}$  for an observed response  $y_{ij}$ . In red, we represent the construction when  $y_{ij}$  is below the LOQ: the probability of being LOQ for an observation at time  $t_{i,j}$  is computed from the predictive distribution as  $p(y_{ij} < LOQ)$ . We then sample a prediction discrepancy  $pd_{ij}^{sim}$  for this observation in the uniform distribution  $U(0, p(y_{ij} < LOQ))$  (31). We can also use the cumulative density function to find the corresponding simulated value within the predictive distribution and impute back the censored data as shown in Fig. 1 (left) to an imputed value  $y_{ij}^{sim}$ , which can then be used as an observation in the evaluation graphs, in a reverse operation from the first case in blue. In the middle plot, we show the pd. corresponding to censored (red) and uncensored (blue) observations within the distribution of pd. over all the observations and in the plot on the right the corresponding npd values after a normal transformation.

Prediction discrepancies have a known theoretical distribution due to their relationship with the cumulative density function of the observations (16). Indeed, if the model holds, their distribution is simply the uniform distribution over  $[0,1]$ . This property can be used to propose tests and graphs. However, repeated measurements introduce a correlation between the npd in an individual which lead to an inflation in the type I error of the tests. To correct this, Brendel *et al.* proposed to take into account the longitudinal nature of the data, using estimated individual variance-covariance matrix obtained from the data simulated under the model to decorrelate both simulated and observed data before computing the quantiles (17), and called the resulting variable prediction distribution errors (pde). Again we transform pde to the normalised prediction distribution errors (npde) to facilitate their interpretation as residuals. The building approach depicted in Fig. 1 applies to both npd and npde, the difference being that with prediction distribution errors the cumulative density function is that of the decorrelated simulations, and we consider the quantiles for the decorrelated observations. Imputation to complete the dataset in the presence of censored observations is performed from the imputed  $pd_{ij}$ , and the decorrelation step then uses the complete dataset including imputed observations.



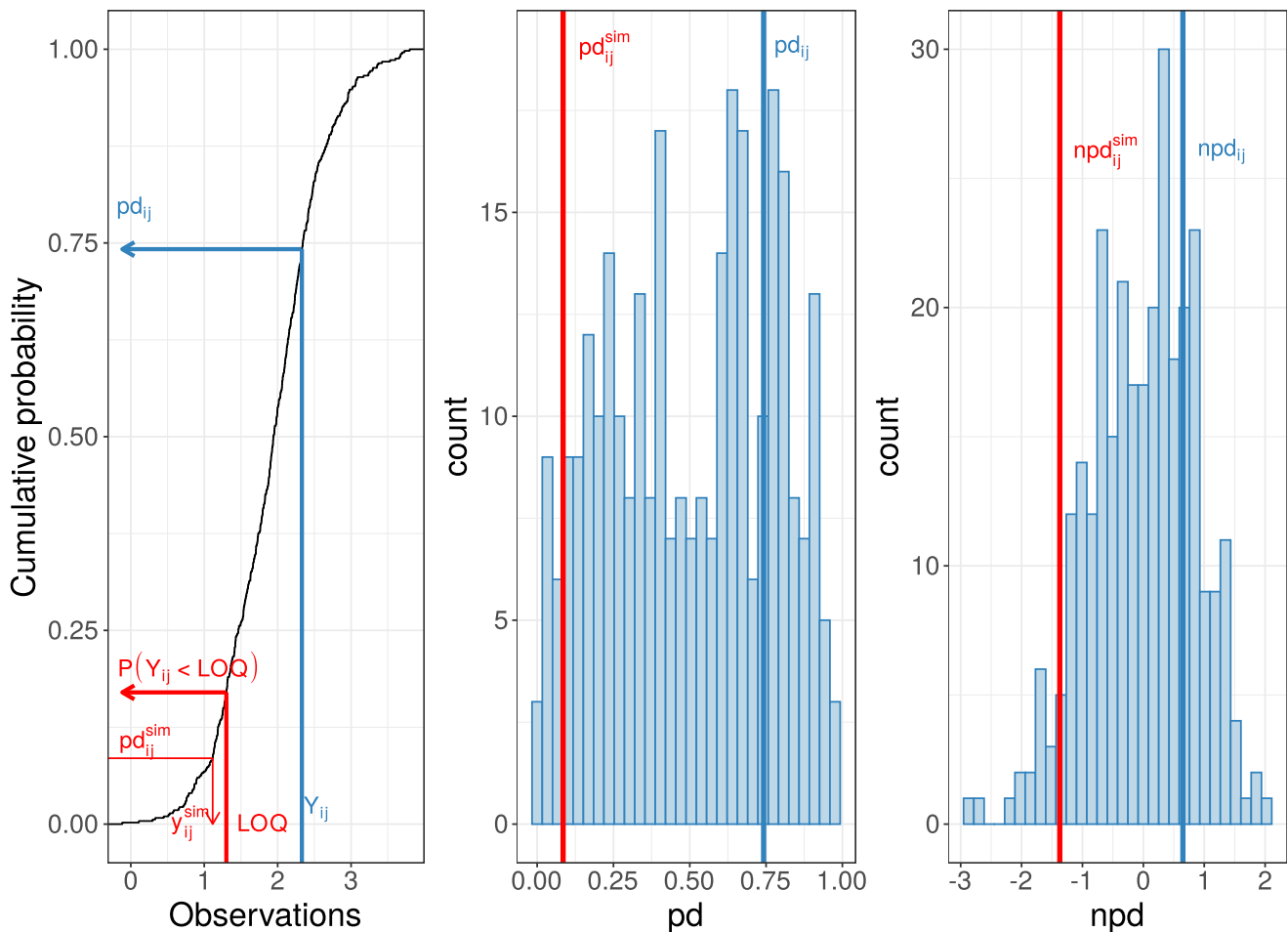


Fig. 1. Construction of the prediction discrepancies

**Graphs and Tests**

Evaluation graphs are the prime way to evaluate models and communicate results. Through their interpretation as residuals, npd and npde can be used to explore the model both for global adequacy through their known distribution and for specific features. Distribution graphs include QQ-plots and histograms, while scatterplots versus the independent variable and versus model predictions can be used to detect when the model deviates from the observed data, suggesting adjustments in the structural or variability models (17). The next section and the appendix show some examples of these diagnostic graphs. A very useful feature in diagnostic graphs is the ability to visualise uncertainty through prediction intervals around selected percentiles of the observed data (34). For comparison with VPC, which show the envelope of simulated responses, npd scatterplots can be transformed by means of a reference profile which takes into account the underlying distribution and within-subject observations (35). Graphs can also be stratified by factors to explore for instance covariate relationships (20). We present an example of this in an [Appendix](#).

Evaluation graphs for simple TTE models are less informative, as outcome and time are confounded, and evaluation relies on distribution plots such as QQ-plots to assess the shape of the distribution. We also suggested de-

trended distribution plots and applied these evaluation tools to joint models, showing how different types of model misspecifications can be detected in departures from the expected predicted intervals on distributions for the npde-TTE (36). In the simulation scenarios evaluated, however, in particular in the absence of random effects associated to the survival model, these graphs provide an evaluation of the overall model which may not be as informative as for continuous models. Further work on repeated TTE models is warranted.

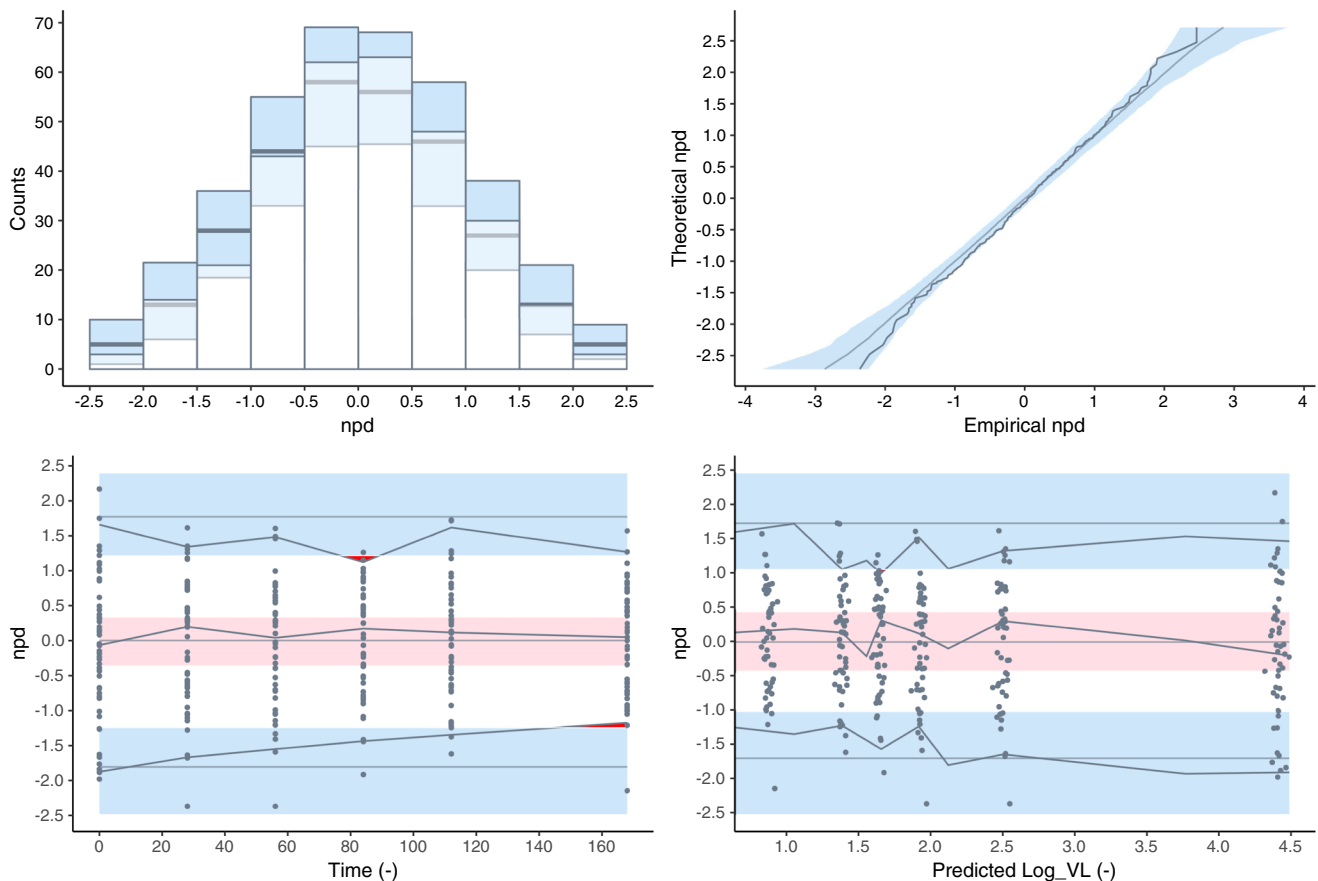
npde is one of the rare metrics to provide a quantifiable assessment of the departure of the observations from the model and can be used to reject candidate models. Since they are known to follow a standard Gaussian distribution, omnibus tests such as the Kolmogorov-Smirnov test, or a combination of three tests evaluating mean, variance and distribution shape, can be used to test them versus their theoretical distribution (17).

**THE NPDE PACKAGE**

In November 2007, the first version of the npde package for R was uploaded to the CRAN Repository (26) and included functions to compute npd and npde as well as fully customisable evaluation graphs. Version 2.0 was released in October 2012 and implemented methods to handle data

below the limit of quantification. Censored data could be omitted, set to a fixed value, imputed to the individual or population prediction or, as we recommend in (29), imputed via the predictive distribution. The current version 3.1, released in February 2021 and updated in April 2021, includes an overhaul of the graphical library to the ggplot2 library (37), diagnostics for covariate models and the possibility to add reference profiles. Figure 2 shows an example of the default evaluation graphs in the npde package, and Fig. 3 presents a screenshot of the output. The graphs were produced for the virload dataset packaged along with the library (the code needed to reproduce these with the npde package is given in an Appendix). This simulated data was based on the phase II clinical trial COPHAR 3-ANRS 134 trial (38), where viral loads after the initiation of an antiviral combination in HIV patients were modelled using a bi-exponential decrease. Fifty patients were simulated, with viral loads measured 6 times over a treatment period of 24 weeks at days 0, 28, 56, 84, 112 and 168. We show the graphs for the complete dataset (without censoring), when the same model is used to compute npde as the one used to simulate the original data. We used 1000 simulations here, a similar number to what is typically used to produce VPC.

The two top graphs in Fig. 2 display information about the distribution of the npd, as a histogram on the left and as a QQ-plot on the right. In both cases, the blue area represents the 90% prediction interval around the distribution. The two plots on the bottom are scatterplots of npd versus the independent variable (left) and the population predictions (right) and help assess whether there are any trends in the model. In these graphs, prediction intervals are produced for the median (in pink), to show deviations for the typical profile, and for the 2.5th and 97.5th percentile of the npd, to a 95% prediction interval (in light blue), to assess whether the interindividual variability is taken into account properly in the model. The prediction intervals are obtained by drawing a large number of samples from  $N(0,1)$  at each time point and computing the 95% prediction interval of the 2.5th, 50th and 97.5th percentile of these samples at each time. Overall, npd are expected to distribute evenly around the abscissa line (defined by  $y=0$ ) and mostly within the interval  $[-1.96; 1.96]$ . In the graphs presented in Fig. 2, all the distributions remain within their prediction intervals, and the scatterplots show that both the median trend and the boundaries of the prediction interval remain within the expected areas, which is what we anticipate given both sets of data have been simulated with the same model. Accordingly, as shown in the



**Fig. 2.** Evaluation graphs using npde package (version 3.1). Top: distribution of the npde shown as a histogram (left) or a QQ-plot (right), with the blue area representing the prediction intervals obtained using simulations under the theoretical  $N(0,1)$  distribution and the dots representing the npde computed for this dataset. Bottom: scatterplots of the npde versus the independent variable (left) and versus the predictions from the model (right). The lines show the evolution of three empirical percentiles (2.5, 50, and 97.5) for the observed data (dark grey) compared to the model predictions (light grey). The pink band corresponds to the prediction interval for the median of the npde (50th percentile) and the blue bands the prediction intervals for the 2.5 and 97.5th percentiles

```

-----
Distribution of npde :
  nb of obs: 300
    mean= 0.03821  (SE= 0.053 )
  variance= 0.8327  (SE= 0.068 )
  skewness= -0.04464
  kurtosis= -0.2207
-----

Statistical tests
  t-test          : 0.469
  Fisher variance test : 0.032 *
  SW test of normality : 0.845
  Global adjusted p-value : 0.0959 .
  ---
  Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.'
-----

```

**Fig. 3.** Screenshot showing the output of the npde package (version 3.1)

screenshot of the output (Fig. 3), the *p* value of the combined test for npde was non-significant. The graphs are produced by default for npd (a change from previous versions of the library where npde was used in graphs, in accordance with the work done with the ISO<sub>P</sub> group in (29)), but the same plots can be produced for pd., pde and npd, with the reference distribution for pd. and pde being U (0, 1) instead of N (0, 1).

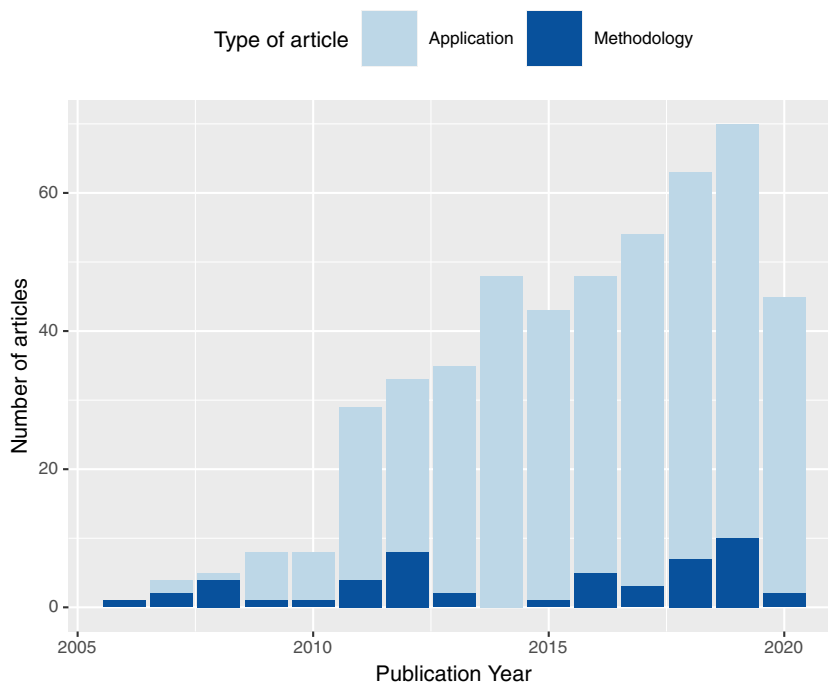
### THE NPDE FOLLOWING

A measure of the popularity in the community is the number of citations generated by an article, and we looked at the citations for the 3 seminal methodological articles (16,17,20), the article presenting the npde package (26) and the tutorial on evaluation metrics (29), which were the most cited on the topic in our bibliography (F. Mentré and E. Comets). We retrieved citations as an Excel sheet from the Web of Science on October 1st, 2020, for each article separately before combining the results in a database. Taken together,

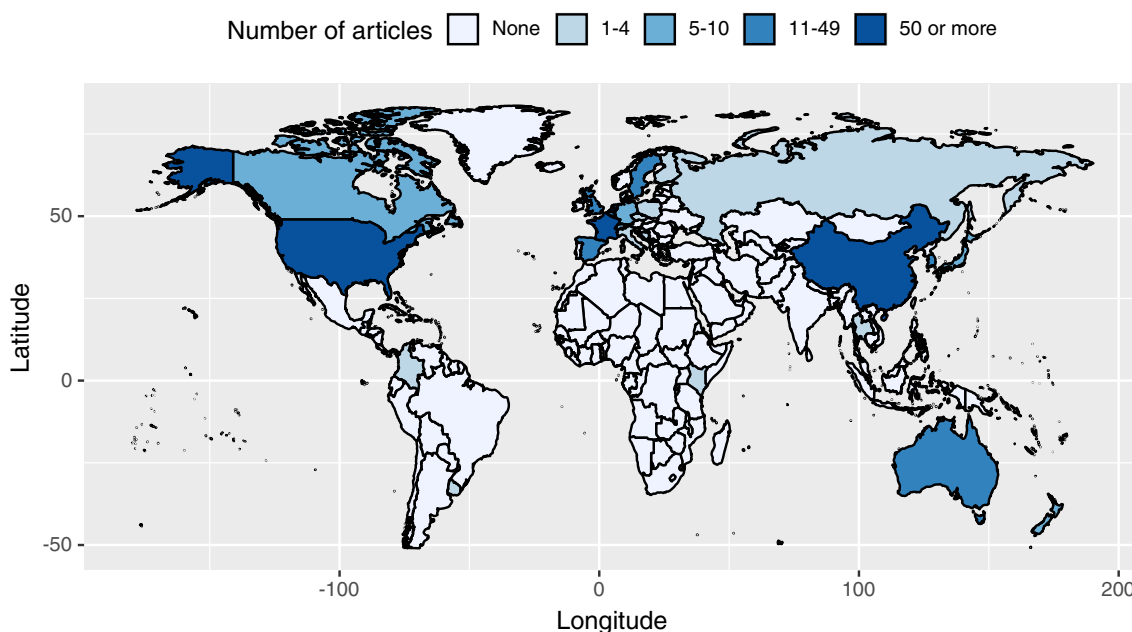
these five foundation articles were cited by 514 unique articles (25 auto-citations), excluding review articles, editorial, abstracts or book chapters (644 citations in total, including 47 auto-citations). Figure 4 shows a histogram of the time of publications of these 514 articles, divided into methodological versus applied based on the title and abstract. The figure shows a large number of articles citations in PK and PK/PD applications after 2011, with a first jump around 30 articles a year and a steady increase up to over 60 articles in 2019, with 3 months left to go in year 2020 at the time of the present article. Considering the more methodological articles, we see several waves corresponding to their birth and evaluation, initial uptake and comparison with other methods and further extensions.

In Fig. 5, we tracked the country given in the address of the corresponding author, producing a world map coloured by the number of publications in each country. The graph shows that our foundation articles have been cited worldwide, with the largest number in France (141 citations), followed by the Netherlands (83), the USA (79) and China (50), and including 2 articles in South America and one in Africa.

Forty methodological articles not authored by one of us cited one of the foundation articles and could be classified in various categories: articles dealing with VPC or their derivatives (*n* = 10), new model evaluation approaches (*n* = 8), estimation methods or algorithms (*n* = 6), modelling approaches (*n* = 10) and tools or guidelines (*n* = 6). The other 463 articles were considered to be more applied, and we looked in a bit more detail at the topics covered, using the keywords reported. Figure 6 shows a word cloud obtained from this data (details concerning data manipulations to extract and format the information plotted in the graph can be found in a notebook in R markdown on Zenodo: <https://zenodo.org/record/4670370>). Applications cover a wide spectrum of terms, some related to the population such as



**Fig. 4.** Number of publications citing one of the five foundation articles, versus year of publication. The articles were classified as methodology or application depending on their main objective. The citations were retrieved from the Web of Science on October 1st, 2020



**Fig. 5.** World map coloured according to the number of articles citing one of the foundation papers. Note that the colours are spaced unequally (none, 1–4, 5–10, 11–49, over 50 articles). The citations were retrieved from the Web of Science on October 1st, 2020

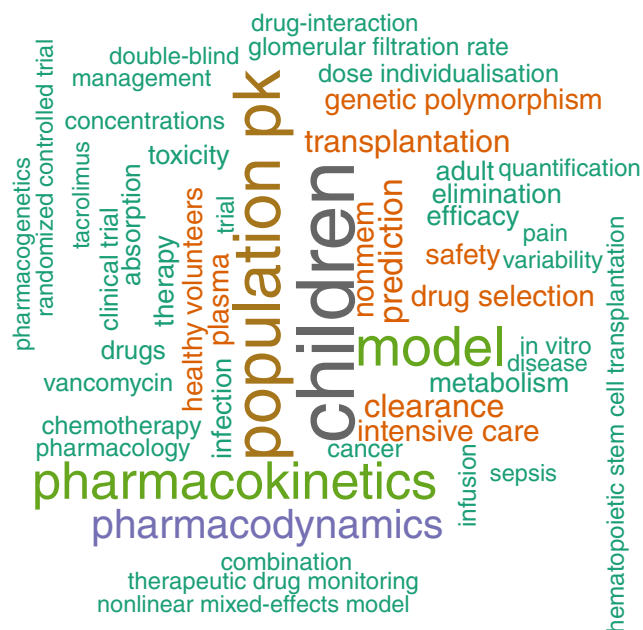
children (regrouping children from newborns to adolescents) or healthy volunteers, some related to the purpose of the analysis such as prediction or therapeutic drug monitoring and others related to the pathology or drug such as intensive care, as well as features of the model.

## CONCLUSION

As NLMEM have taken on an increasingly prominent role in drug development and administration, so has the importance of rigorous evaluation and transparent reporting. The ISO-P tutorial, written by a group of experts in NLMEM and their evaluation, put forth a core set of recommended evaluation graphs for models with continuous outcomes, weighing the pros and cons of each (29). npd and npde are amongst the simulation-based tools recommended by the group. As such, they provide overall diagnostics of the model and behave better than prediction-based diagnostics in situations of poor information leading to shrinkage. On the other hand, they require simulations under the model and an adequate modelling, for instance, through joint models in cases including dropout. More recently, the computation of npd has been extended successfully to censored and discrete data.

While npde have been shown to perform well in simulation studies, in practice we find the statistical tests to be sensitive to outliers which may not be relevant for the purpose of modelling in pharmacometrics, with a sensitivity that increases with the number of observations as the prediction intervals become increasingly narrow. Another issue is the number of simulations to perform, which should be increased with large number of observations to avoid ties in the npde distribution. While tests perform well in simulation studies, in real life even minute differences between reality and the simplified representation provided by a model may cause tests to be significant with large datasets. Also, as Box amusingly puts it, “since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad” (39).

Evaluation graphs are therefore the recommended approach to guide model building and communicate on the performances of alternative models in order to select the best model for a given purpose. This is also the main reason behind the progressive shift of the term “validation” to the more accurate words of “evaluation” or “qualification” as stated in (13). npd and npde can be used both in internal evaluation, using the data used to build the model, and in external evaluation on a separate dataset, which is considered a more robust assessment of the model’s predictive ability.



**Fig. 6.** Word cloud of the most frequent keywords in applied articles citing one of the foundation articles.



In the light of the development of adequate evaluation tools and of the update in guidelines, it would be interesting to evaluate whether the reporting of population PK/PD studies has improved. In 2007, we performed a survey on PK/PD analyses reported in the two preceding years and found that only a fourth of articles included any kind of model evaluation (39,40). As noted in (41) in a review calling for improved repeatability of PK analyses, involving better reporting amongst other open science practices, this survey has not been repeated since.

## SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at <https://doi.org/10.1208/s12248-021-00597-7>.

## AUTHOR CONTRIBUTION

The two (female) authors of this article (F. Mentré and E. Comets) have been the main architects for the development of npde. The original idea of prediction discrepancies was proposed by France Mentré with her PhD student Florence Mesnil for non-parametric mixed effect models and extended to general parametric models with Sylvie Escolano. Emmanuelle Comets got involved in the definition of npde and co-supervised the PhD of Karl Brendel, who also conducted in collaboration with Céline Dartois a survey of how population PK and PD analyses were evaluated. The extension of npde to BLQ data was the methodological topic of Tram (Thi Huyen) Nguyen, another female PhD student supervised by France Mentré. Tram was also the first author for the collaborative white paper by the ISoP group (29). Emmanuelle Comets supervised the PhD of Marc Cerou, our second male PhD student on the topic, who extended npd to models involving time-to-event and categorical outcomes (manuscript under preparation). Finally, we also would like to acknowledge the contribution since January 2020 of our engineer, Romain Leroux, for the latest version of the npde library. France Mentré was awarded the prestigious Lewis Sheiner lecturer award from the University of California and the International Society of Pharmacometrics in 2013, recalling in her lecture the prominent part of model evaluation in her career (2).

## REFERENCES

1. Sheiner LB, Rosenberg B, Marathe VV. Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. *J Pharmacokin Biopharm.* 1977;5(5):445–79.
2. Mentré F. Lewis Sheiner ISoP/UCSF lecturer award: from drug use to statistical models and vice versa. *CPT Pharmacometrics Syst Pharmacol.* 2014;3:e154.
3. Beal S, Sheiner L, Boeckmann A, Bauer R. NONMEM Version 7.2. Ellicott City; 1989-2011.
4. Lavielle M. Mixed effects models for the population approach: models, tasks, methods and tools. Chapman & Hall/CRC Biostatistics Series; 2014.
5. Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. *J Stat Softw.* 2017;76:1–32.
6. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: linear and nonlinear mixed effects models; 2020. R package version 3.1–150. Available from: <https://CRAN.R-project.org/package=nlme>.
7. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw.* 2015;67(1):1–48.
8. Comets E, Lavenu A, Lavielle M. Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. *J Stat Softw.* 2017;80:1–41.
9. Fidler M, Xiong Y, Schoemaker R, Wilkins J, Trame M, Hooijmaijers R, et al. nlmixr: nonlinear mixed effects models in population pharmacokinetics and pharmacodynamics; 2021. R package version 2.0.1. Available from: <https://CRAN.R-project.org/package=nlmixr>.
10. Mentré F, Mallet A, Baccar D. Optimal design in random-effects regression models. *Biometrika.* 1984;84:429–42.
11. Sheiner LB, Rosenberg B, Melmon KL. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput Biomed Res.* 1972;5(5):441–59.
12. Sheiner LB, Beal SL. Some suggestions for measuring predictive performance. *J Pharmacokin Biopharm.* 1981;9:503–12.
13. European Medicines Agency. Guideline on reporting the results of population pharmacokinetic analysis (CHMP); 2007. Available from: [http://www.ema.europa.eu/docs/en\\_GB/document\\_library/Scientific\\_guideline/2009/09/WC500003067.pdf](http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003067.pdf).
14. Food and Drug Administration. Guidance for industry exposure-response relationships—study design, data analysis, and regulatory applications; 2019. Available from: <https://www.fda.gov/media/128793/download>.
15. Karlsson M, Savic R. Diagnosing model diagnostics. *Clin Pharmacol Ther.* 2007;82:17–20.
16. Mentré F, Escolano S. Prediction discrepancies for the evaluation of nonlinear mixed-effects models. *J Pharmacokin Pharmacodyn.* 2006;33:345–67.
17. 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;23:2036–49.
18. Mesnil F, Mentré F, Dubruc C, Thénot JP, Mallet A. Population pharmacokinetic analysis of mizolastine and validation from sparse data on patients using the nonparametric maximum likelihood method. *J Pharmacokin Pharmacodyn.* 1998;26(2):133–61.
19. Dunn PK, Smyth GK. Randomized quantile residuals. *J Comput Graph Stat.* 1996;5:236–44.
20. Brendel K, Comets E, Laffont C, Mentré F. Evaluation of different tests based on observations for external model evaluation of population analyses. *J Pharmacokin Pharmacodyn.* 2010;37:49–65.
21. Yano Y, Beal SL, Sheiner LB. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. *J Pharmacokin Pharmacodyn.* 2001;28(2):171–92.
22. Holford N. The visual predictive check—superiority to standard diagnostic (Rorschach) plots. *PAGE* 14. 2005;Abstr 738.
23. Bergstrand M, Hooker A, Wallin J, Karlsson M. Prediction-corrected visual predictive checks for diagnosing nonlinear mixed-effects models. *AAPS J.* 2011;13:143–51.
24. Hooker AC, Staats CE, Karlsson MO. Conditional weighted residuals (CWRES): a model diagnostic for the FOCE method. *Pharm Res.* 2007;24(12):2187–97.
25. Nyberg J, Bauer RJ, Hooker AC. Investigations of the weighted residuals in NONMEM 7. *PAGE* 10. 2010;Abstr 1883.
26. 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 Meth Prog Biomed.* 2008;90:154–66.
27. R Core Team. R: a language and environment for statistical computing. Vienna, Austria; 2015. Available from: <https://www.R-project.org/>.
28. Keizer R, Karlsson M, Hooker A. Modeling and simulation workflow for NONMEM: tutorial on Pirana, PsN, and Xpose. *CPT Pharmacometrics Syst Pharmacol.* 2013;2:e50.
29. Nguyen T, Mouksassi MS, Holford N, Al-Huniti N, Freedman I, Hooker A, et al. Model evaluation of continuous data pharmacometric models: metrics and graphics. *CPT Pharmacometrics Syst Pharmacol.* 2017;6(2):87–109.

30. Bergstrand M, Karlsson M. Handling data below the limit of quantification in mixed effect models. *AAPS J.* 2009;11(2):371–80.
31. Nguyen THT, Comets E, Mentré F. Extension of NPDE for evaluation of nonlinear mixed effect models in presence of data below the quantification limit with applications to HIV dynamic model. *J Pharmacokinet Pharmacodyn.* 2012;39(5):499–518.
32. Cerou M, Lavielle M, Brendel K, Chenel M, Comets E. Development and performance of npde for the evaluation of time-to-event models. *Pharm Res.* 2018;35(2):30.
33. Mallet A. A maximum likelihood estimation method for random coefficient regression models. *Biometrika.* 1986;73:645–56.
34. Comets E, Brendel K, Mentré F. Model evaluation in nonlinear mixed effect models, with applications to pharmacokinetics. *J Soc Fr Statistique.* 2010;151:106–28.
35. Comets E, Nguyen THT, Mentré F. Additional features and graphs in the new npde library for R. *PAGE* 22. 2013;Abstr 2775.
36. Cerou M, Peigné S, Chenel M, Comets E. Performance of npde for the evaluation of joint model with time to event data. *PAGE* 28. 2019;Abstr 8940.
37. Wickham H. *ggplot2: elegant graphics for data analysis.* New York: Springer-Verlag; 2016. Available from: <https://ggplot2.tidyverse.org>
38. Savic R, Barrail-Tran A, Duval X, Nembot G, Panhard X, Descamps D, et al. Effect of adherence as measured by MEMS, ritonavir boosting, and CYP3A5 genotype on atazanavir pharmacokinetics in treatment-naive HIV-infected patients. *Clin Pharmacol Ther.* 2012;92:575–83. [39] box G. science and statistics. *J Am Stat Assoc.* 1976;71:791–9.
39. Brendel K, Dartois C, Comets E, Lemmenuel-Diot A, Laveille C, Tranchand B, et al. Are population PK and/or PD models adequately evaluated? A 2002 to 2004 literature survey. *Clin Pharmacokin.* 2007;46:221–34.
40. Dartois C, Brendel K, Comets E, Laffont C, Laveille C, Tranchand B, et al. Overview of model building strategies in population PK/PD analyses: 2002 to 2004 literature survey. *Br J Clin Pharmacol.* 2007;64:603–12.
41. Ioannidis J. Reproducible pharmacokinetics. *J Pharmacokinet Pharmacodyn.* 2019;46:111–6.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.