Title: Model evaluation of continuous data pharmacometric models: metrics and graphics

Authors: Thi-Huyen-Tram Nguyen (1), Mohamad-Samer Mouksassi (2), Nick Holford (3), Nidal Al-Huniti (4), Immanuel Freedman (5), Andrew C. Hooker (6), Jyothy John (7), Mats O. Karlsson (6), Diane R. Mould (8), Juan José Pérez Ruixo (9), Elodie L. Plan (10), Rada Savic (11), Johan G.C. van Hasselt (12), Benjamin Weber (13), Chenguang Zhou (14), Emmanuelle Comets (1,15) and France Mentré (1) for the Model Evaluation Group of the International Society of Pharmacometrics (ISoP) Best Practice Committee.

Affiliations:

(1) INSERM, IAME, UMR 1137, F-75018 Paris, France; Université Paris Diderot, Sorbonne Paris Cité, Paris, France

(2) Certara Strategic Consulting, Montréal, Canada

(3) Department of Pharmacology & Clinical Pharmacology, University of Auckland, Auckland, New Zealand

(4) Quantitative Clinical Pharmacology, AstraZeneca, Waltham, Massachusetts

(5) Dr Immanuel Freedman Inc., Harleysville , PA USA

(6) Dept. of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

(7) Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Washington, USA

(8) Projections Research Inc, Phoenixville, PA USA

(9) The Janssen Pharmaceutical Companies of Johnson & Johnson, Belgium

(10) Pharmetheus, Uppsala, Sweden

(11) Dept. of Bioengineering and Therapeutic Sciences, UCSF, San Francisco, USA

(12) Division of Pharmacology, Leiden Academic Centre for Drug Research, Leiden University, Leiden, Netherlands

(13) Boehringer Ingelheim Pharmaceuticals, Inc., CT, USA

(14) Genentech, San Francisco, USA

(15) INSERM CIC 1414, Rennes, France; University Rennes-1, Rennes, France.

Running title: Evaluation graphs for population PKPD models

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1002/psp4.12161

Corresponding author:

France Mentré, 16 rue Henri Huchard, 75018 Paris, France, france.mentre@inserm.fr

Key words: model evaluation, diagnostics, graphs, pharmacometrics, nonlinear mixed effect model, pharmacokinetic, pharmacodynamic

Word count: 7761

Figure count: 7

Table count: 1

Reference count: 36

INTRODUCTION

(75 words)

This article represents the first in a series of tutorials on model evaluation in nonlinear mixed effect models (NLMEM), from the ISoP Model Evaluation Group. Numerous tools are available for evaluation of NLMEM, with a particular emphasis on visual assessment. This first basic tutorial focuses on presenting graphical evaluation tools of NLMEM for continuous data. It illustrates graphs for correct or misspecified models, discusses their pros and cons and recalls the definition of metrics used.

BACKGROUND

Nonlinear mixed effects models (NLMEM) have been established as the state of the art methodology in pharmacometrics for analysis of longitudinal pharmacokinetic (PK) and pharmacodynamics (PD) measurements collected in preclinical and clinical studies, especially in drug development[1–3]. NLMEM for continuous PKPD data use nonlinear dynamic models that draw on physiological or pharmacological principles to provide a reasonable approximation of the dynamics of the drug in the body and of their effects. They describe both population and subject specific characteristics, represented as fixed parameters for population characteristics and random parameters for subjects. NLMEM are widely applied because of their ability to quantify several levels of variability, to handle unbalanced data, and to identify for individual-specific covariates.

In any regression modelling, after fitting a model to a dataset, it is essential to assess the goodness-of-fit between the model and the dataset and to determine whether the underlying model assumptions appear appropriate. In this tutorial, we refer to this procedure as 'model evaluation' although in literature, it has been described under several more or less equivalent terms such as 'model diagnostics', 'model adequacy', 'model assessment', 'model checking', 'model appropriateness' and 'model validation'. Model evaluation has to be clearly distinguished from 'model building' and 'model qualification' processes, which are two steps of model development that require model evaluation but imply different concepts. 'Model building' is the process of developing a model on a given dataset to achieve clearly defined analysis objectives. 'Model qualification' is the assessment of the performance of a model in fulfilling the analysis objectives. 'Model evaluation' is required for both processes to diagnose one or several intermediary or key models in a model building step or evaluate a selected model with respect to the modeling objectives. In this tutorial, we will only focus on the model evaluation step and describe various tools for evaluation of a NLMEM, regardless of whether it is an intermediary model, a key model in model building step or a best model that can be used for further inferences.

Although there are many statistical tools for model evaluation, the primary tool for most biomedical science and engineering modeling applications is graphical analysis. Graphical methods have an advantage over numerical methods for model evaluation because they readily shine light on a broad range of complex aspects of the relationship between the model and the data. Different types of graphical analyses evaluating a fitted model provide information on the adequacy of different aspects of the model. NLMEM methodology is naturally linked with many assumptions related to executed design (e.g. unbalance design), data collection, form of structural model, multiple levels of variability to be quantified, residual model, and covariate model. Interactions between model components such that misspecification of one component may have consequences for the apparent appropriateness of other components in the fitted NLMEM adds to the challenge of model evaluation, therefore a large set of tools is required.

In recent years, many new methods for graphical model evaluation have been developed. For example, new residual-based model diagnostics have been developed (e.g., Conditional Weighted Residuals (CWRES), Normalized Prediction Distribution Errors (npde), etc.),[4–6] new models for the study of variance have been proposed, shortcomings of commonly used Empirical Bayes Estimates (EBE) have been underlined[7] and methods for using visual predictive checks have been developed and described.[8,9]

To provide a compass and fit-for-purpose direction in the emerging and very active field of model-based drug development, ISoP Best Practice Committee has initiated a 'Model Evaluation Group' to provide detailed guidance for model evaluation of NLMEM. This basic tutorial represents the first in a series of tutorials on model evaluation. It describes a core set of graphical tools for evaluation of NLMEM for continuous data and provides guidance, especially to beginner modelers, on how they are meant to be used. It includes a description of the different metrics, an illustration about their graphical use on 'true and 'misspecified' models, and a discussion of their pros and cons. Each metric is first described by equations and then by a less technical explanation in order to make the definition of each tool easier to understand for readers with statistical or pharmacometric backgrounds. The graphs are broadly separated into two categories: prediction-based (basic tools) and simulation-based (Table 1). The use of each graph is illustrated via two case examples, which were chosen to illustrate the properties and behaviors of different

evaluation tools in different situations. This is not necessarily to mimic what should be done in real-world modeling. Therefore, in these two examples we used only simulated data and ignored important steps such as exploring data, researching literature to understand the data and to propose a set of plausible models, etc. The first example is a contrived example, in which we used a very simple PK model. The designs (dose, distribution of covariates, and allocation of sampling times) were selected to easily show the properties of various evaluation tools although they may not resemble real data conditions. In the second part of the tutorial, we applied the presented graphical tools to evaluate a more complex example that is based on a real study. This is a pharmacokinetic pharmacodynamic (PKPD) model describing the total warfarin concentration and its effect on prothrombin complex activity. For each example, we simulated data from a 'true' model and fitted different models including the 'true' and various 'misspecified' models to highlight some types of model deficiency. To show the availability of several software tools developed for NLMEM estimation and simulation, the steps of simulating data, estimating parameters and computing evaluation metrics were performed using a variety of software, including R (https://cran.r-project.org/), NONMEM (http://www.iconplc.com/innovation/solutions/nonmem/), MONOLIX (http://lixoft.com/products/monolix/), and PHOENIX (https://www.certara.com/software/pkpd-modeling-and-simulation/phoenix-nlme/). The graphical displays presented in the tutorial are a suggestion using R scripts but it is not a recommendation from the ISOP Model Evaluation group.

PHARMACOKINETIC CASE EXAMPLE

Structural model

A simple PK model for a hypothetical drug was utilized throughout sections 3 and 4 to illustrate different evaluation tools for detecting various types of model misspecification. The PK model is a two-compartment model with first-order elimination following a single IV bolus administration. Time course of drug concentration is described by Eq. 1:

$$\frac{dA_1}{dt} = -\frac{CL}{V_1} \times A_1 - \frac{Q}{V_1} \times A_1 + \frac{Q}{V_2} \times A_2 \quad (1)$$

$$\frac{dA_2}{dt} = \frac{Q}{V_1} \times A_1 - \frac{Q}{V_2} \times A_2$$

$$C(t) = \frac{A_1}{V_1}$$

The model has four parameters CL, V_1, Q, V_2 representing clearance, volume of distribution for the central compartment, inter-compartment clearance and volume of distribution for the peripheral compartment, respectively. We evaluated a binary covariate effect of concomitant treatment (without = 0, with = 1) on clearance and a linear effect of body weight on the central volume of distribution. Half of the hypothetical patients received a concomitant treatment in addition to the drug. Body weight is assumed to be distributed normally across the population, with a mean value of 70 kg and standard deviation of 15 kg.

Statistical model

The central compartment drug concentration y_{ij} for individual i, observed at time t_{ij} is given by:

$$y_{ij} = f(\theta_i, t_{ij}) + \varepsilon_{ij} \quad (2)$$

where f is the PK function, which is identical for all the individuals; θ_i is a vector of p individual parameters for the individual i and ε_{ij} is the residual error. Let y_i denote the vector of n_i observations y_{ij}, t_i the vector of n_i sampling times t_{ij}, ε_i the vector of n_i residual errors, while ε_{ij} with j=1…n_i. ε_i is assumed to follow a normal distribution with mean 0 and variance $\Sigma(\theta_i, t_i)$ as follows:

$$\Sigma(\theta_i, t_i) = \left(\sigma_{add} + \sigma_{prop} f(\theta_i, t_i)\right)\left(\sigma_{add} + \sigma_{prop} f(\theta_i, t_i)\right)' \quad (3)$$

For this example, we assumed a combined error model with σ_{add}, $\sigma_{prop} \neq 0$.

The vector of individual parameters θ_i can be characterized by a function h of the fixed effects, representing the typical population values of the parameters and random effects η_i, specific for each individual. The random effects are assumed to follow a multinormal distribution with mean 0 and variance Ω, N(0, Ω). Here we assumed that each PK parameter follows a log-normal distribution, therefore h has an exponential form. For instance, for the k^{th} parameter, h is given by:

$$\theta_{ik} = h(\mu, \eta_{ik}) = \mu \times \exp(\eta_{ik}) \quad (4)$$

The variance-covariance matrix Ω of the random effects is assumed in this PK model to have all diagonal elements equal to ω^2, where ω is the standard deviation of the individual random effects for each PK parameter. Of course this is not a common situation since the variability of parameter usually differs from each other. We also assumed a correlation between the random effects of CL and V_1, i.e., a non-null covariance term between CL and V1 in the matrix Ω. A part of the inter-individual variability may be explained by including covariates. In the presence of covariates, the individual parameters θ_{ik} are described by:

$$\theta_{ik} = h(\mu, \eta_{ik}, z_i) = \mu \times \exp(\beta \times z_i) \times \exp(\eta_{ik}) \quad (5)$$

where z_i is the vector of covariates for individual i, which can be a binary (or categorical) covariate (e.g., concomitant treatment in this example) or a continuous covariate (e.g. body weight in this example) and β is the vector of covariate effect. Of note, a transformation was made for the body weight so that the reference profile corresponds to that of a patient with a weight of 70 kg.

$$T_{\text{Weight\_i}} = \log\left(\frac{\text{Weight}_i}{70}\right) \quad (6)$$

We call Ψ the vector of population parameters, where $\Psi = \{\mu, \Omega, \beta, \sigma_{add}, \sigma_{prop}\}$.

Study design and data simulation

We considered a balanced design, with 5 sampling times per patient (t = 0.5, 1, 4, 12, 24h after treatment) which we call the standard design. Two other designs, referred to as the sparse and very sparse designs, with two or one sampling times per patient, respectively, were also used to illustrate the influence of shrinkage on some of the evaluation graphs. In these sparse sampling designs, samples were randomly selected from the five sampling times above. A dataset of 180 patients with three treatment groups each receiving one of the three different doses of 10, 100 and 1000 mg, was simulated using the model with the sampling designs described above and parameters provided in **Supplementary Table S1**.

Evaluated scenarios

To illustrate the properties of various evaluation graphs in different situation where the model may or may not be appropriate for describing the data, we fitted several models to the simulated data: i) the true model which was used for data simulation, ii) a misspecified structural model, in which the structural model was changed into a one-compartment model, iii) a model without covariates, in which no covariate effect was considered, iv) a misspecified correlation model, in which we neglected the correlation between the random effects of clearance and distribution volume, and v) two misspecified residual error models, in which we considered only a constant ($\sigma_{add} \neq 0$, $\sigma_{prop} = 0$) or proportional error model ($\sigma_{add} = 0$, $\sigma_{prop} \neq 0$). Note that we modified only one property of the true model at the same time to obtain these different types of model misspecification.

Parameter estimation, computation of evaluation metrics and software

Population parameters were estimated using the default option of the Stochastic Approximation Expectation Maximization algorithm implemented in MONOLIX 4.3.3(http://lixoft.com/products/monolix/). The simulated data and medians of predictions for different models are shown in **Supplementary Figure S1**.

Individual parameter estimates were then estimated as Empirical Bayes Estimates (EBE), that is, as the mode of the posterior distribution, conditional on the observed data and the population model. Let $p(\eta_i|y_i,\Psi)$ the conditional distribution of η_i. The EBE estimate of η_i is given by:

$$\hat{\eta}_i = \text{argmax}_{\eta_i}\big(p(\eta_i|y_i,\Psi)\big) = \text{argmax}_{\eta_i}\left(\frac{p(y_i|\eta_i,\Psi)\times p(\eta_i|\Psi)}{p(y_i)}\right) \quad (7)$$

At this stage, μ has been estimated. Therefore once $\hat{\eta}_i$ is estimated, $\hat{\theta}_i = h(\mu,\hat{\eta}_i,z_i)$ can be easily calculated.

Once the model parameters were estimated, evaluation metrics were computed using additional software. In this paper, all the goodness-of-fit graphs were generated using R and several R packages for which scripts available in **Supplementary R Codes**. The full process for generating evaluation graphs for the different models is summarized in **Supplementary Figure S2**.

BASIC EVALUATION TOOLS

Evaluation based on population predictions

Population predictions

By definition, population predictions (xPRED, with x = ∅, C or P) are the expectation of the model, $E(y_i)$, given the individual designs and covariates. There are several methods for computing xPRED. The simplest way is model linearization using the First-Order linearization (FO) (Eq. 8) , i.e. the prediction assuming all random effects equal 0 (denoted PRED). The corresponding predicted profile, given design and covariates, is often call the prediction for typical individual, or typical profile.

$$E(y_i) \approx PRED_i = f(h(\mu, 0, z_i), t_i), \tag{8}$$

An alternative method is to use First-Order Conditional Expectation (FOCE) approximation giving predictions denoted CPRED (Eq. 9):

$$E(y_i) \approx CPRED_i = f(h(\mu, \hat{\eta}_i, z_i), t_i) - \frac{df(h(\mu,\eta_i,z_i),t_i)}{d\eta_i}\Big|_{\eta_i=\hat{\eta}_i} \hat{\eta}_i{}', \tag{9}$$

Another method for computing xPRED is to use Monte Carlo simulation, in which, population predictions (denoted PPRED) are defined as the mean of the model predictions (Eq. 10). By definition, PPRED is the 'average' prediction (response) of the population.

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

where $y_i^{sim(k)}$ is the vector of data obtained at the k^{th} simulation using the model and the design of the individual i (t_i). As in NLMEM, $E(f(h(\mu,\eta_i),t_i) \neq f(\mu,t_i)$, the 'average prediction' for the population (PPRED) in general differs from PRED, the prediction for a 'typical' patient, especially for models with high inter-individual variability and high nonlinearity.

Observations can be plotted versus population predictions to evaluate the population model (the first and second rows of **Figure 1A**). The line of identity (and sometimes a local regression line) is added to the graph. Even if the model is correctly specified, the data points are not necessarily scattered around the line of identity but the regression line will be more or less close to the identity line. This can be seen in the first column of the first and second rows of **Figure 1A**, where the data were fit to the true model. A systematic departure of the data points or the trend line from the identity line (as seen in the first column of the first and second rows of **Figure 1A**) could indicate a misspecification in the structural model. On the other hand, misspecification in the residual error model is difficult to detect using these graphs (i.e., the first column of the third and last rows of **Figure 1A**) because the residual error model is not considered in the computation of xPRED. Trends of deviations between the local regression and identity lines could appear independently of model misspecification because in NLMEM, the observations are not symmetrically distributed around the mean. This can occur especially for models with high nonlinearity and large inter-individual variability, regardless of the method of xPRED computationare computed. Trends can also appear by using linearization to compute xPRED or by neglecting the intra-individual correlation or the heterogeneity of residual errors or the presence of data below the quantification limits when calculating the local regression line.[7] It is also useful to examine the graph of observations versus population predictions in both normal and log-scale in order to better evaluate the quality of fit, especially when the data cover several orders of magnitude.

Population residuals

The population residuals (– xRES with x = ∅, C or P) are defined as the difference between the observations and population predictions ($xRES_i = y_i - xPRED_i$). These residuals are correlated within each individual and their magnitude may depend on that of observations if the residual error model is not homogeneous (i.e., an additive error model), which we call heteroscedastic. Population weighted residuals (xWRES) standardize and decorrelate the population residuals using the model-predicted variance-covariance matrix of observations, $Var(y_i)$:

$$xWRES_i = Var(y_i)^{-\frac{1}{2}} \times (y_i - xPRED_i) \quad (11)$$

Depending on the methods used to compute population xPRED and model-predicted Var(y_i), there are various types of population weighted residuals. The classical population weighted residuals, termed WRES, are calculated from PRED and Var(y_i) obtained with the FO approximation (Eq. 8 & Eq.12).

$$Var(y_i) \approx \left.\frac{df(h(\mu,\eta_i,z_i),t_i)}{d\eta_i}\right|_{\eta_i=0} \Omega \left.\frac{df(h(\mu,\eta_i,z_i),t_i)'}{d\eta_i}\right|_{\eta_i=0} + \Sigma(h(\mu,0,z_i),t_i) \quad (12)$$

Another type of population weighted residuals, termed Conditional weighted residuals (CWRES), is obtained from CPRED and Var(y_i) computed by FOCE (Eq. 9 & Eq.13).

$$Var(y_i) \approx \left.\frac{df(h(\mu,\eta_i,z_i),t_i)}{d\eta_i}\right|_{\eta_i=\hat{\eta}_i} \Omega \left.\frac{df(h(\mu,\eta_i,z_i),t_i)'}{d\eta_i}\right|_{\eta_i=\hat{\eta}_i} + \Sigma(h(\mu,\hat{\eta}_i,z_i),t_i) \quad (13)$$

Weighted residuals can also be obtained using PPRED and Var(y_i) calculated from Monte Carlo simulation (Eq. 10 & Eq. 14). These residuals are called Population or Expectation weighted residuals (PWRES or EWRES) in MONOLIX and NONMEM, respectively.

$$Var(y_i) \approx \frac{1}{K}\sum_{k=1}^{K}(y_i^{sim(k)} - E(y_i))(y_i^{sim(k)} - E(y_i))' \quad (14)$$

Among these three types of population weighted residuals, WRES was shown to result in misleading diagnoses in some instances, especially when the model becomes highly non-linear, which causes the FO approximation to be poor.[5,7] CWRES obtained by FOCE and PWRES obtained by Monte Carlo simulation have been shown to have better performance in evaluation of NLMEM[5,7] and we present graphs for these types of weighted residuals for the PK example. By definition, if the model is true, xWRES should have zero mean and unit variance. However, unlike weighted residuals of linear mixed models which should follow a normal distribution if the model is correct, the xWRES of NLMEM have an unknown distribution because the marginal distribution of observations is not normal.

Various graphs based on population weighted residuals have been proposed to evaluate NLMEM, such as the scatterplots of population weighted residuals versus time (first and second rows of **Figure 1B**) or versus population predictions (first and second rows of **Figure 1C**). If the model is true, the population weighted residuals should be randomly scattered around the horizontal zero-line as shown in the first row of **Figure 1B** and **1C**), where the true model was fitted to the data. A systematic bias from the zero-line may imply deficiencies in the structural model (second row of **Figure 1B** and **1C**) but can also be a consequence of informative censoring or adaptive designs. A misclassified error model can be identified from the amplitude of the residual distribution along the x-axis, e.g., a cone-shape pattern of residuals would suggest a heteroscedastic error model (third and last rows of **Figure 1B** and **1C**. Trends that appear when conditioning on covariates (first and second rows of **Figure 1D**) or when plotting population weighted residuals versus covariates may suggest a problem in the covariate model or of the need to include covariates in the model. Of note, as in NLMEM, observations may not be distributed symmetrically around the mean, the population weighted residuals, regardless of how they are computed, are not necessarily distributed evenly around the horizontal zero line even in absence of misspecification, especially for models with high nonlinearity with respect to random effects and large inter-individual variability. Another point to bear in mind when examining these weighted residuals is that decorrelation using the full variance-covariance matrix may cause some modifications in the trend lines, for instance, the position where the trend appears in the plots versus time or predictions may be different from where it is when examining graphs of normal residuals (RES) versus time or versus predictions.

Evaluation based on individual predictions and individual random effects (EBE)

Individual predictions and individual residuals

Individual estimated vector of random effect ($\hat{\eta}_i$), i.e. EBEs, can be used to calculate other individual-based evaluation metrics such as individual predictions, IPRED and individual weighted residuals, IWRES.

$$IPRED_i = f(h(\mu, \hat{\eta}_i, z_i), t_i) \quad (15)$$

$$IWRES_i = \Sigma(h(\mu, \hat{\eta}_i, z_i), t_i)^{-\frac{1}{2}} \times (y_i - IPRED_i) \quad (16)$$

Several types of graphs based on these individual metrics can be used for model evaluation. First of all, the graph of IPRED versus observations offers a global assessment of the individual fit for all patients, mainly to identify a misspecification in structural model (last row of **Figure 1A**). The considerations provided for the observations vs population predictions graph are also applicable here. Deficiencies in structural and residual error models can be detected using the scatterplot of IWRES versus time (last row of **Figure 1B**) or individual predictions (last row of **Figure 1C**). The graphs based on individual predictions or residuals are similar to those based on population predictions but with less variability because inter-individual variability was taken into account in their computation. Therefore, in some cases, model misspecification can be detected more easily with individual-based metrics. However, unlike population-based metrics, individual predictions and residuals do not allow for evaluation of covariate models (last row of **Figure 1D**) as the variability that results from any existing covariates that is not taken into account will be considered to be part of inter-individual variability and therefore, is included in the estimated individual random effect, $\hat{\eta}_i$.

Finally, the individual fit, obtained by superposing the individual observations and the individual predictions over the independent variable in the same graph, is one of the most frequently presented evaluation graphs. It provides a simple way to visualize whether the model is able to describe individual data profiles. A substantial discordance between predictions and observations could indicate a problem in the population model, either in the structural model or the variability model. However it would be difficult to determine the primary cause based solely on this graph (**Supplementary Figure S3**). This graph is also useful for identifying practical problems in the data such as sample switching, bioanalysis errors, etc. However, with a large number of patients, it will be challenging to examine all the individual fitted graphs. Population predictions can be added to the individual fits to provide more information, for instance, about shrinkage (see below) or how an individual response differs from that of the population. This is a helpful way of seeing the individual predictions appear to be randomly scattered around the population prediction.

Evaluation based on Empirical Bayes estimates

The estimated EBEs can also be used as an evaluation metric. They can be used to evaluate the inter-individual variability model. For each component of the vector of EBEs (or of the vector individual parameter estimates) graphs such as a histogram or a boxplot could be drawn, and compared to their estimated predicted population-distribution. A substantial discordance between an EBE distribution and a population distribution may imply misspecification of the random effect models. For instance, a multimodal distribution of an EBE would suggest the need to include covariates or use a parameter distribution other than the assumed distribution (e.g., log-normal distribution) for the corresponding random effect; a high kurtosis in an EBE distribution may indicate poor individual information or an incorrect underlying distribution assumption and suggest that variability of this random effect may be poorly estimated. EBEs can also be plotted versus each other to identify correlation between random effects (**Supplementary Figure S4A**). High correlation between EBEs of several parameters may also indicates a problem in model parameterization (over-parameterization or non-identifiability of model parameters). If the model correctly handles random effect correlation, one expects to see almost no trend in the graphs with decorrelated EBE s, $\hat{\eta}_i^*$, (**Supplementary Figure S4B**), obtained by standardizing the EBEs using the estimated variance-covariance matrix of random effects $\widehat{\Omega}$ (Eq. 17).

$$\hat{\eta}_i^* = \widehat{\Omega}^{-\frac{1}{2}} \times \hat{\eta}_i \tag{17}$$

EBE-based evaluation graphs can also be used to detect deficiencies in the structural model. For instance, the graph of each EBE component versus doses is one of the simplest methods to detect model deficiencies for drugs with nonlinear PK. Another important and frequent use of EBE-based evaluation graphs is to screen covariate effect and evaluate a covariate model. This can be done by examining the correlation between EBE component and a continuous covariate or a boxplot stratified by different classes of a categorical covariate. If the model correctly takes into account the covariate effect, a correlation would be expected between the covariate and the corresponding individual parameter estimates (**Supplementary Figure S5A**), but no correlation should remain between the corresponding EBE and the covariate (**Supplementary Figure S5B**). With rich individual information, the square value of the coefficient of

correlation between the covariate and the individual parameter estimates represents the fraction of inter-individual variability that is explained by the covariate.

Influence of shrinkage on individual-based evaluation tools

The estimation of EBE and individual predictions is susceptible to a phenomenon called shrinkage that occurs when the individual data are not sufficiently informative with respect to one or more parameters.[10] Under these conditions, the EBE and individual parameter estimates would shrink close to the population mean. This phenomenon can be quantified by η-shrinkage, estimated by 1- sd(EBE)/ω or 1- var(EBE)/ω², where ω is the inter-individual standard deviation estimated in the population model.[10] The fact that individual predictions may tend to the individual observations for more or less sparse designs can be quantified by ε-shrinkage, defined as 1-sd(IWRES) or 1-var(IWRES), where IWRES are the individual weighted residuals.[10] Of note, there are two definitions of shrinkage in literature, one based on the ratio of variances and one based on the ratio of standard deviations.

With high shrinkage, the individual-based evaluation tools become less informative and do not allow for a correct evaluation of a model. For instance, a high η-shrinkage may hide or falsely induce the true relationships or distort the shape of the EBE distribution, of the correlation between EBEs (**Supplementary Figure S6A**) or of the correlation between EBEs and covariates (**Supplementary Figure S6B**).[7,10] If overfitting occurs, IWRES will shrink towards 0, which makes model evaluation based on this metric less effective or informative.[7] **Supplementary Figure S7A** provides IPRED vs observations plots, at different levels of ε-shrinkage, of the misspecified structural model. **Supplementary Figure S7B** shows the same graphs of population predictions in which we can also see the influence of limited information.

SIMULATION-BASED EVALUATION TOOLS

Simulation-based evaluation tools were first developed by Bayesian statisticians and are now increasingly used to evaluate NLMEM. They rely on the concept of the Posterior Predictive Check (PPC)[11], whose principle is that if a model describes correctly a dataset, the data simulated under that model would be similar to the observations. Hence, to evaluate a model with these methods, one needs to simulate a large

number (K) of Monte Carlo samples under the tested model, using the design of the observed dataset then compare a statistic computed from the observed data with that computed from the simulated data. The chosen statistic can be a pharmacokinetic parameter calculated from non-compartmental analysis, for instance, area under the curve AUC, half-life, steady state concentration, maximum or minimum concentration,[12,13] or other statistical inferences such as mean prediction errors (residuals), root mean square prediction errors[11,13] or objective function value.[14]

Instead of using a statistic that condenses all the information of the observations into a single value, one can also evaluate a model using all the observations, by comparing the observations with their predicted distribution. These observations-based PPCs such as the visual predictive check (VPC),[8] prediction discrepancies (pd), and normalized prediction distribution errors (npde)[4,6] are among the most frequently used simulation-based evaluation tools. We may compare the observed statistics or observations with their distribution via graphical assessment or statistical tests. The use of statistical tests is not considered in this tutorial but is discussed in several papers.[6,11,15–17]

Visual and Numerical Predictive Check

The VPC offers a graphical comparison of the distribution of observations and the distribution of predictions versus an independent variable such as time, dose or other covariates.[8]. It comprises in comparing the distribution of the observations with that of the predictions using different percentiles of the distributions. A classical presentation of VPC, originally termed “scatter VPC”, is obtained by plotting the observations together with the predicted percentiles of the simulated data (usually 2.5, 50, 97.5th or 5, 50, 95th or 10, 50, 90th percentiles) over the independent variable.[8,9] The area defined by the lowest and highest percentiles is usually called the prediction interval (PI) of the data, for instance, 10 and 90th percentile define a 80% PI (**Supplementary Figure S8A**). The “scatter VPC” characterizes model appropriateness by comparing the number of observations included within or outside a selected PI of data with a theoretical value. For instance, in a VPC with a 80% PI, we expect to have 80% of the observations within the 80% PI, 50% above and 50% below the median, 10% above the 90th percentile and 10% below the 10th percentile.

A more visually intuitive representation of VPC has been proposed and has rapidly gained popularity within the last few years.[9] In this new version of VPC, termed "Confidence interval VPC", the percentiles of the observations, the predicted percentiles and their 95% confidence interval are plotted (**Supplementary Figure S8B**).[9] For this type of VPC, data binning, in which an independent variable (usually time) has to be split into several bins, each containing an approximately equal number of observations, is necessary to calculate the percentiles of observed and predicted data.[9] The confidence intervals (CI) of the predicted percentiles are obtained using the K Monte Carlo samples: the same selected percentiles, e.g., 2.5, 50, 97.5th percentiles, are calculated for each of the K simulated datasets. The 95% CI for each of the selected percentiles is then easily obtained from the distribution of the K percentiles computed for the K simulated datasets. If the model is correct, the observed percentiles should be close to the predicted percentiles and remain within the corresponding CI. However, an appropriate VPC may still result from models where individual unexplained variability is misspecified, either because it is absorbed by other model components (inter-occasion variability by residual variability for instance) or because it is contributing little to the overall variability (e.g., residual variability in the presence of large inter-individual variability).

An important property of VPC is that it conserves the original units (e.g., time, concentration, etc.) of the model, therefore appears familiar. However, VPC also has some drawbacks. First of all, data binning (frequently necessary to construct a "Confidence interval VPC") is challenging for unbalanced designs with differing number of observations at each time points and may influence the interpretation of VPC.[18,19] Secondly, heterogeneity in design such as differing doses, dosing regimen, route of administration or covariates may render the VPC for the whole data non-informative.[20] For instance, in **Supplementary Figures S8A-B**, the upper, median and lower predicted percentiles or CI correspond to the observations following the highest, median and lowest dose, respectively. In such cases, data stratification by dose and/or by important covariates or dose normalization may mitigate these problems. However, data stratification often leads to a loss of power and dose normalization may not be appropriate for nonlinear pharmacokinetics (i.e., the model is nonlinear with respect to dose). Prediction-corrected VPC (pcVPC)

offers a solution to these problems while retaining the visual presentation of the VPC (**Supplementary Figure S8C-D**).[20] In a pcVPC, the variability in each bin is removed by normalizing the observed and simulated dependent variables based on the typical population prediction for the median independent variables.[20] The observation y_{ij} and the corresponding simulated data $y_{ij}^{sim(k)}$ are corrected by the given formulae:

$$_{pc}y_{ij} = y_{ij} \times \frac{xPR\tilde{E}D_{bin}}{xPRED_{ij}} \quad (18)$$

$$_{pc}y_{ij}^{sim(k)} = y_{ij}^{sim(k)} \times \frac{xPR\tilde{E}D_{bin}}{xPRED_{ij}} \quad (19)$$

where $xPR\tilde{E}D_{bin}$ is the population prediction for the median independent variables in a specific bin and $xPRED_{ij}$ is the population prediction for individual i at time j. In the seminal paper, Bergstrand et al proposed to use PRED for population predictions[20] while in the PK example, we calculated the pcVPC using PPRED. Of note, for continuous covariates, a VPC may be constructed using a covariate, such as body weight or age or model predictions, as the independent variable. This does not lose power like data stratification methods and can provide useful confirmation of the appropriateness of a covariate model.[9,21]

Despite these limitations, the VPC offers a very intuitive assessment of misspecification in structural, variability and covariate models therefore has now become a widely used evaluation tool for evaluating NLMEM. **Figure 2A-B** show the VPC plots of different models. We can clearly see that observed percentiles remain within the corresponding intervals for the true model (first column) while clear departure from the confidence interval is evident for misspecified structural, residual error (**Figure 2A**) and covariate models (**Figure 2B**).

The numerical version of VPC, known as the Numerical Predictive Check (NPC), is also used in model evaluation.[9,22] It summarizes the information of several "scatter VPC" evaluated at different selected PIs, for instance, the 0, 20, 40, 50, 60, 80, 90, 95% PI.[9,22] NPC calculates the percentages of outliers for each selected PIs, which are the observed data above and below different PIs. By providing the same calculation

for each of the K simulated datasets, we can obtain a CI for the percentages of outliers. The observed percentages can be compared with the empirical CI using a coverage plot. Just as in a VPC plot, a trend in NPC coverage plot would indicate a misspecification of the structural, inter-individual variability or residual error model (**Supplementary Figure S9**). As NPC evaluates model misspecification on several PIs, it may provide additional information compared to the VPC, which only presents one selected PI. Also, it compares each observation with its own simulated distribution, so normalization and stratification to handle the binning as in the VPC is not necessary. However, unlike VPC, which is a representation of observations and predictions versus time, NPC loses the time dimension, therefore, would not be able to point out at which time points the model over- or under-predicted the data.

Prediction discrepancies & normalized prediction distribution errors

Prediction discrepancies are a form of observation-based PPC, developed for NLMEM by Mentré and Escolano.[4] Let F_{ij} denote the cumulative distribution function of the observation y_{ij} for the individual i. The prediction discrepancy of this observation is defined as its percentile in the predictive distribution, given by:

$$pd_{ij} = F_{ij}(y_{ij}) = \int^{y_{ij}} p(y|\Psi)\, dy = \int^{y_{ij}} \int p(y|\theta_i, \Psi) p(\theta_i|\Psi) d\theta_i\, dy \tag{20}$$

Using the predictive distribution approximated by the K Monte Carlo simulation samples, the prediction discrepancy is then calculated by:

$$pd_{ij} = F_{ij}(y_{ij}) = \frac{1}{K}\sum_{k=1}^{K} 1_{y_{ij}^{sim(k)} < y_{ij}} \tag{21}$$

The computation of the pd uses the same simulations as the VPC. The pd of an observation y_{ij} are indeed computed as the number of times that the simulations are under the observation in a VPC.

A decorrelated version of pd, the prediction distribution error (pde), has been proposed in order to enable the use of statistical tests.[6] The pde are computed in the same way of pd but from decorrelated (observed and simulated) data, obtained using Eq. 11 with PPRED and $Var(y_i)$ calculated by a MC method (Eq. 10 & 14):

$$pde_{ij} = F_{ij}^{*}(y_{ij}^{*}) = \frac{1}{K}\sum_{k=1}^{K} 1_{y_{ij}^{sim(k)*} < y_{ij}^{*}} \quad (22)$$

where y_{ij}^{*} and $y_{ij}^{sim(k)*}$ are decorrelated observed and simulated data, respectively. By construction, pd and pde are expected to follow a uniform distribution of zero-mean and unit variance, U[0,1] if the model describes the data adequately. The normalized pd and pde, denoted npd and npde, calculated by Eq.23, follow a normal distribution with zero-mean and unit variance, N(0,1):

$$\begin{aligned} npd_{ij} &= \Phi^{-1}(pd_{ij}) \\ npde_{ij} &= \Phi^{-1}(pde_{ij}) \end{aligned} \quad (23)$$

where Φ is the inverse function of the cumulative distribution function of N(0,1).

As the npd and npde naturally account for the heterogeneity in study design by comparing the observations with their own distribution, no data stratification or dose normalization is required for a global evaluation using these metrics, although covariate model exploration can benefit from stratifying the npde plots versus covariates values or categories (**Supplementary Figure S10A**). This is an advantage of npd or npde compared to the traditional VPC. The assessment of npd and npde could be done using several types of graph such as scatterplots versus time or predictions, quantile-quantile (q-q) plots or histograms, scatterplots or boxplots versus continuous or categorical covariates or doses. For those who prefer the presentation of a VPC, where the original units of the model and data are conserved, a transformed version of the npd and npde has been proposed to take into account the shape of data evolution over time (**Supplementary Figure S10B**).[23] As in a VPC, the observed percentiles and CI of predicted percentiles can also be added into the evaluation graphs of npd and npde, which requires binning the data. Like graphs based on population residuals, scatterplots of npd or npde versus time or predictions are helpful to detect and distinguish different types of model misspecification, e.g. structural, residuals or covariate. **Figure 3A-C** show the graphs of npd versus time and predictions for different models. We can see that misspecification in structural model, error model and covariate model can be detected by the departure of the observed percentiles from their prediction intervals. For graphical evaluation, it may be sometimes better to use npd

instead of npde because decorrelation can induce artifacts, i.e. create trends or make trends less apparent in the scatterplots of npde versus time or predictions,[17] as we can see in **Figure 3D-E**.

CORE SET OF COMMON GRAPHS FOR MODEL EVALUATION

Because NLMEM relies on several assumptions and that no evaluation tool can address all the components of a model, a comprehensive evaluation of a pharmacometric model would usually require examination of several diagnostic graphs as described in detail in the previous section. A brief description of each evaluation tool and which model components it addresses is summarized in **Table 1** and **Supplementary Table S3**.

In this tutorial, we recommend a core set of common graphs that are useful in most situations for comprehensive evaluation of a pharmacometric model (column "In core set" in **Table 1**). This core set of evaluation graphs includes basic prediction-based graphs such as population or individual predictions versus observations, individual fits, EBE correlation graphs and simulation-based graphs (VPC and npd). However, any diagnostic based on EBEs (i.e., EBE, IPRED, IWRES), may be misleading in the presence of substantial shrinkage. For population predictions and residuals, we recommend to use CPRED (and CWRES) or PPRED (and PWRES), depending on estimation method. More specifically CPRED and CWRES could be used for FOCE estimation, and PPRED and PWRES when the estimation methods do not involve linearization.

For VPC, we recommend to use VPC showing observed and predicted percentiles such as the 'Confidence Interval' VPC or the 'Percentile' VPC[9]. The scatter VPC is discouraged because it is hard to compare the observation distribution with the predictions especially when there are many data points. The pcVPC is recommended if there are important covariate effects or different dose groups or an adaptive trial design has been used.

For EBEs correlation graphs, scatterplots of decorrelated EBEs can also be examined if the variance-covariance matrix of random effects was not diagonal.

To evaluate a covariate model, some additional graphs are required according to the column "core set of graphs" in **Table 1.**

PHARMACOKINETIC PHARMACODYNAMIC CASE EXAMPLE

In this section, we illustrate the use of the core set of evaluation graphs on a PKPD example that is more realistic for modeling practice. In this example, we only evaluate the PD model and as there is no covariate effect in the PD model, no graphs for evaluating covariate model is displayed. **Structural and statistical models**

The exemplary PKPD model for warfarin is based on data reported by O'Reilly et al.[24,25] The PK model, describing total warfarin concentration following a single dose administration, is a one-compartment model with first order absorption, a lag-time and first-order elimination. A turnover PD model with an inhibitory E_{max} function on the rate of Prothrombin Complex Activity (PCA) production describes the effect of warfarin on PCA. The structural model for the PKPD of warfarin is described by the following set of differential equations:

$$\frac{dA_1(t)}{dt} = \begin{cases} 0 & if\ t < T_{lag} \\ -\frac{\ln(2)}{T_{abs}} A_1(t) & if\ t \geq T_{lag} \end{cases}$$

$$\frac{dA_2(t)}{dt} = \frac{\ln(2)}{T_{abs}} A_1(t) - \frac{CL}{V} A_2(t)$$

$$C(t) = \frac{A_2(t)}{V}$$

$$\frac{dPCA(t)}{dt} = R_{in}\left(1 - \frac{E_{max} C(t)}{C(t) + C_{50}}\right) - \frac{ln(2)}{T_{eq}} PCA(t) \quad (24)$$

$$A_1(0) = 0$$

$$A_2(0) = 0$$

$$PCA(0) = \frac{R_{in}}{\left(\frac{ln(2)}{T_{eq}}\right)}$$

where $A_1(t)$ and $A_2(t)$ are the warfarin amounts at the site of absorption and in the central compartment, respectively. C(t) is the total warfarin concentration, CL, V, T_{abs} and T_{lag} represent clearance, central volume of distribution, absorption half-life and lag time, respectively. CL and V were allometrically scaled by body weight using 70 kg as the reference value with the allometric exponent β fixed at 3/4 and 1, respectively (Eq. 25). This is an equivalent expression of Eq.5 with the transformation described in Eq.5&6. There are no covariate effects on the PD parameters. PCA(t) denotes the activity of prothrombin complex. PCA is produced with a zero-order rate R_{in} and eliminated with first-order rate constant k_{out}, equal to $\ln(2)/T_{eq}$, where T_{eq} is the half-life of PCA elimination. PCA was assumed to be at steady state before administration of warfarin with the baseline $PCA_0 = R_{in}/k_{out}$. The models were parameterized using PCA_0 instead of R_{in}. E_{max} is a parameter denoting the maximum possible effect of warfarin and C_{50} denotes the concentration of warfarin that results in half maximal inhibition.

$$\theta_i = h(\mu, \eta_i, z_i) = \mu \times \left(\frac{Weight}{70}\right)^{\beta} \times \exp(\eta_i) \quad (25)$$

We assumed log-normal distribution for CL, V, T_{abs} T_{lag}, E_{max}, C_{50}, PCA_0 and T_{eq} (Eq. 4). A combined additive and proportional residual error model was used for the PK predictions and an additive residual error model for the PD predictions. Model parameters are provided in **Supplementary Table S2**.

Study design and data simulation

We considered the same design as the one used in the original study.[24] There were 27 male and 5 female patients (N=32). Body weight ranged from 40 to 102 kg. The administered doses were calculated on a 1.5 mg per kg basis. Central compartment drug concentrations (mg/L) and PCA (PCA unit) were measured as described in the original report. A dataset was simulated using the design and models described above.

Evaluated scenarios

To mimic different types of model misspecification, we fitted several models to the simulated data: a) a misspecified structural PD model with effect immediately related to concentration (Misspecified delay, Immediate effect), b) a misspecified structural PD model with an effect compartment (Misspecified delay, Effect compartment), c) a misspecified structural PD model and variability model, in which an effect compartment was used for the PD part and covariance between all PK parameters in one block and all PD parameters in another block was considered (Misspecified delay and correlation, Effect compartment, Full Omega) d) the true model which was used for data simulation (True model, Turnover PD).

Parameter estimation, computation of evaluation metrics and software

Data simulation was performed using NONMEM version 7.3 (http://www.iconplc.com/innovation/solutions/nonmem/). For parameter estimation, the PK model was first fitted to the simulated concentrations. The population PK parameters were then fixed at their estimated values to perform data fitting for the PCA observations. Thus, only the PD model was misspecified in this example. Parameter estimation was performed using the Quasi-random Parametric Expectation Maximization (QRPEM) algorithm[26] in Phoenix NLME 7.0

(https://www.certara.com/software/pkpd-modeling-and-simulation/phoenix-nlme/) keeping all default settings. For simulation-based evaluation graphs, a thousand replicates were performed.

Core set of graphs for model evaluation

In this section we go through a core set of model evaluation graphs, for each PKPD model (**Figure 4** for the Misspecified delay, Immediate effect model, **Figure 5** for the Misspecified delay, Effect compartment model, **Figure 6** for the Misspecified delay and correlation model and **Figure 7** for the true turn-over model) and provide our assessments in a more general sense to illustrate how these evaluation graphs may be interpreted if the true model is not known. Here as our model does not contain covariates, we did not present the specific graphs that can be used to evaluate covariate model. For the basic prediction-based graphs, we presented PPRED and PWRES graphs since we estimated parameters using Expectation-Maximization algorithm.

Misspecified delay, immediate effect model

Because the data covers a large range, although the scatterplots of observations versus PPRED or IPRED or the graphs of PWRES or IWRES versus predictions are not very informative as there is a big gap in the data, the trends in those graphs can still show that there is a discordance between the data and model (**Figure 4A**, first and last columns). Splitting those graphs into several graphs with different scales may allow for easier interpretation. Clear trends can be observed in the graphs of PWRES or IWRES versus time, showing that the structural model is not sufficient to describe the observations (**Figure 4A**, second column). The discordance between the model and observations can also be seen in the individual fit graphs of three randomly selected individuals (**Figure 4B**) as well as in the simulation-based graphs (**Figure 4D)**. Very high correlations between all PD parameters may be a result of misspecification of the structural model (**Figure 4C**).

Misspecified delay, Effect compartment model

The basic (prediction-based) goodness-of-fit plots (**Figure 5A**) as well as the individual fits (**Figure 5B**) do not show any important disagreement between the data and model predictions. However, the

simulation-based graphs with confidence intervals added show that the chosen structural model can describe the median but not sufficiently characterize the upper and lower percentiles (**Figure 5D**). High correlations between all PD parameters may indicate that the model is misspecified (**Figure 5C**).

Misspecified delay and correlation model

Similar to the previous model, the basic goodness of fit plots (**Figure 6A**) as well as the individual fit plots (**Figure 6B**) do not allow to see an important discordance between the data and model. The simulation-based diagnostics show that the model describes well the median but seems to fail to capture the 90% percentile and the 10% percentile.

True model – Turn over model

All the evaluation graphs of the turnover model shows that the model seems to be adequate to describe the data (**Figure 7**) and reveals no problem with the parametrization of the model (**Figure 7C**). Some potential underestimation of the upper 90% percentile can be observed in the VPC and npd vs time plots and this could indicate that the model may still have some problems in describing the variability. However, in this case, it results from the limited number of observations or patients included in the analysis (**Figure 7D**) because the model is known to be correctly specified. This illustrates why the percentiles and confidence interval should not be used too rigorously to either accept a model or identify problems.

For completeness, **Supplementary FigureS12** presents the NPC coverage for different models. The immediate effect model had some Observed/Expected ratios outside of the CI. Even the Observed/Expected ratios of the two effect compartment models that remained within the CI show a systemic departure from the expected ratio of 1, indicating that the models may not be appropriate. Finally, an NPC coverage plot of the turn over model with the lower and upper ratio close to 1 confirms that this model seems sufficient to describe the observed data.

DISCUSSION

Numerous model evaluation metrics have been developed to evaluate NLMEM and their underlying assumptions. As mentioned earlier, we focus here only on graphical tools used in model evaluation and not

in model building nor in model qualification even though model evaluation is involved in the two latter steps of modeling. A brief description of each evaluation tool and which model components it addresses is provided in **Table 1** and **Supplementary Table S3**.

Besides the graphical tools presented in this paper, numerical tools and methods such as model identifiability assessment,[27] parameter estimation standard errors[2,12], resampling-based methods such as the bootstrap[28] or model selection criteria (-2 log likelihood or objective function, AIC, BIC)[2,12,29,30] are always used in parallel to provide additional information for the reliability of a model or for model comparison. However, in this first tutorial about model evaluation, we chose to focus only on graphical tools for several reasons. Firsty, graphical tools are commonly used and reported tools for model evaluation. Secondly, unlike numerical tools which compact all the information of model misspecification in a single statistic such as a p-value, graphical tools can show how much the model is able to characterize the data, where it fails to do so and therefore provide hints for model misspecification. Finally, many numerical tools (e.g. objective function, AIC, BIC) are only useful for comparisons between models (in model building or qualification) and do not allow the evaluation of a single model for its adequacy. Of course, we cannot neglect their role in model development.

The evaluation metrics and their visualization discussed in this paper can be classified in two categories: prediction-based and simulation-based metrics. Prediction-based metrics, except for PPRED and PWRES, two metrics that are computed by Monte Carlo simulation, can be provided with little computational burden and therefore, can be used to assess model appropriateness in every step of model building. However, population prediction-based metrics computed using FO or FOCE linearization may result in misleading evaluation when the linearization used is inappropriate. Individual prediction-based metrics may be not sufficiently reliable at high level of shrinkage (see specific section for more details). For this reason, shrinkage should be evaluated and reported to provide information about relevance of the individual prediction-based evaluation tools. Currently, there is no consensus on the level of shrinkage which renders these individual metrics no longer reliable. A shrinkage value of 30% or 50%, if calculated from standard deviation or variance, respectively, has been suggested as a threshold for high shrinkage[7,10]

but whether this threshold should be applied for all models and population parameter values remains to be evaluated.

As indicated by their name, simulation-based metrics requires data simulation. However, many design and data features such as adaptive design, response-guided treatment, changes in dosing regimen to limit adverse effect or to maintain drug concentrations within the therapeutic window (therapeutic drug monitoring), missing data, or drop-out, etc. may be difficult to reproduce through simulation. If these features are neglected, simulation-based graphs may show significant trends even though the underlying model is adequate to describe the data.[20,31,32] One solution is to develop joint models that describe the longitudinal data and the features in question, for instance a time-to-event coupled with longitudinal data to handle drop-outs so that the link between the two processes may be correctly taken into account during simulation.[31–33]As a large number of simulations is required (usually ≥ 1000 simulations),[34] the computation of these metrics can be a time-consuming process, especially with complex models and large data sets, and therefore, may not be always feasible for evaluating models when the timelines for decision making are tight, especially in an industrial setting. However, the resource limitations related to time may become less severe with the widespread availability of high performance computing and better project management.

In general, the uncertainty in model parameters is not always accounted for when computing simulation-based evaluation metrics. Using a simple PKPD example, Yano et al found that using point estimates provided similar results to other approaches which consider not only the point estimates but also parameter uncertainty. Nevertheless, they did mention that this conclusion may depend on study designs and the extent of interindividual variability.[11] Samtani et al suggested that parameter uncertainty could be ignored in data simulation if it is negligible with respect to the between and within subject variability and the sample size available.[35] Otherwise, the computation of simulation-based metrics should account for uncertainty in the model parameter.[35] Even though simulation-based evaluation tools now have become standard evaluation graphs which are reported in most population analysis, they may still remain less familiar than classical evaluation tools such as predictions or weighted residuals to audiences composed of non-modelers.

Throughout the two examples, we show the properties of several evaluation tools in various situations where we have different types of model misspecification. Of note, in this tutorial, we employed at many instances the terms 'model misspecification' or 'model deficiency' as in both examples, we used simulated data and fitted them to different models which were already known to be true or false. In our opinion, these two terms should not be used in real world data driven analysis as there is no 'true' model and other terms such as 'goodness-of-fit' or 'agreement with data' are more appropriate.

Various tools have been developed and are required to evaluate NLMEM because no single evaluation tools or planar graphs can effectively address all aspects of the model components. To detect a specific misspecification or to identify a problem, one diagnostic graph may be sufficient (**Table 1** and **Supplementary Table S3)**. However, a model may present deficiencies in various components and a misspecification of one component may conceal misspecification in other components. Therefore, in our opinion, in order to comprehensively evaluate a model, the core set of graphs proposed in this article should be examined.

In summary, we presented in this tutorial different evaluation tools and illustrated their graphical use to evaluate simple pharmacometric models that may arise during all the processes of model development. We also defined a core set of common graphs that may be shown and examined during model evaluation. As the target audience of this tutorial is beginner modelers with statistical or pharmacometric background, we did not attempt to provide an exhaustive list of all the existing evaluation tools and methods for NLMEM but restricted ourselves to some of the most frequently used tools in pharmacometrics. Although some methods have been proposed to account for several factors that can influence model evaluation as mentioned in the previous paragraph such as adaptive design,[20] data below the limit of quantification[36] or drop-out,[32,33] we avoided describing these more advanced methods and chose to keep this subject for a future tutorial of the ISoP Model Evaluation group. We also focused only on model evaluation for continuous longitudinal data, which represents a significant portion of population modeling. Model evaluation for other types of data and models such as discrete data, categorical data, time-to-event data, etc. requires additional evaluation tools and would merit additional tutorials.

ACKNOWLEDGMENTS

This work was performed by members of the Model Evaluation (MoEv) Group of the "ISoP Best Practice Committee". The authors would like to thank other members of the group that participated in the outline of this first tutorial: Malidi Ahamadi, Brian Corrigan, Kevin Dykstra, Marc Lavielle, Robert Leary, Don Mager, Peter Milligan, Flora Musuamba Tshinanu, Rune Overgaard, Marc Pfister, Richard Upton and Byon Wonkyung. The authors would like also to thank René Bruno, chair of the ISoP Best Practice Committee, for his involvement in the preparation of the tutorial.

TABLE LEGEND

Table 1. Various evaluation graphs in NLMEM[(1)] and proposal for a core set of evaluation graphs

FIGURE LEGENDS

Figure 1. Basic goodness-of-fit plots for true model (first column), misspecified structural model (second column), misspecified constant error model (third column) and misspecified proportional error model (last column). (A) Observation versus population predictions calculated using FOCE method (CPRED, first row), MC simulation (PPRED, second row) or individual predictions (IPRED, third row). Identity and local regression lines are presented in black and red, respectively. This graph clearly points out that the wrong structural model under-predicted high concentrations while misspecification of the residual error model is more difficult to be detected using this type of goodness-of-fit plots. **(B)** Weighted residuals (CWRES, PWRES, IWRES from the first row to third row, respectively) versus time plots. xWRES are shown as blue points, spline lines are also added in these graphs as the red curves. A systematic trend indicates a misspecification in the structural model (second column). A cone-shape of residuals indicates a problem of residual errors (third and last columns). (C) Systemic biases when conditioning on covariates reveal a deficiency in the covariate model. **(C)** Weighted residuals (CWRES, PWRES, IWRES from the first row to third row, respectively) versus population predictions plots. **(D)** Weighted residuals (CWRES, PWRES, IWRES from

the first row to third row, respectively) versus time, stratifying on binary covariate (concomitant treatment yes/no). Systemic biases in the population weighted residuals versus time plots when conditioning on covariate indicates the need to include the covariate in the model (third and second lines). Unlike population weighted residuals, no visually significant trend can be found in the plots of IWRES versus time of the model lacking covariates. This emphasizes once again that IWRES-based graphs cannot be used to evaluate a covariate model.

Figure 2. pcVPC plots of the true model (first column), misspecified structural model (second column), misspecified constant error model (third column) and misspecified proportional error model (last column). The blue and red lines are the observed percentiles (10, 50, 90th percentiles), the blue and red ribbons are the corresponding 95% confidence intervals. The dashed black lines are predicted percentiles. Observations corresponding to the lowest, median and highest doses are shown in blue, pink and green, respectively. (A) A systemic departure of the observed percentiles from the prediction intervals could indicate a misspecification in structural or residual error model (B) Trends observed when stratifying on covariates helps to evaluate the covariate model.

Figure 3. Scatterplots of npd or npde versus time or population predictions (PPRED) for the true model (first column), misspecified structural model (second column), misspecified constant error model (third column) and misspecified proportional error model (last column). The blue and red lines are the observed percentiles (10, 50, 90th percentiles), the blue and red ribbons are the corresponding 95% confidence intervals. The dashed black lines are predicted percentiles. Observations corresponding to the lowest, median and highest doses are shown in blue, pink and green, respectively. **(A)** npd versus time (first row). A systematic trend indicates a misspecification in the structural model. A cone-shape of residuals indicates a problem of residual errors. **(B)** npd versus PPRED (second row). **(C)** npd versus time stratified by binary covariate for the true (first and second columns) and misspecified covariate model (third and last columns).

Systemic biases when conditioning on covariate reveal a deficiency in the covariate model. **(D)** npde versus time. In this case, decorrelation makes the trends become less apparent, especially for the misspecified proportional error model. **(E)** npde versus time for the true (first and second columns) and misspecified covariate model (third and last columns), stratified by the binary covariate (concomitant treatment). In this case, decorrelation makes the trends to detect a lack of covariate effect become less apparent.

Figure 4. Core set of diagnostic graphs for Misspecified delay, Immediate effect model. (A) Basic goodness-of-fit plots. Blue points denotes individual data points. Lowess lines are in red. Interpretation should focus on areas where the data are available with little attention to gaps. **(B)** Representative randomly selected individual fits. Blue points denotes individual observed data points, green curve is the IPRED, red curve is the PRED. **(C)** Correlations and histogram of Empirical Bayes Estimates. **(D)** Simulation-Based Diagnostic Plots (VPC, npd). The blue and red lines are the observed percentiles (10, 50, 90^{th} percentiles), the blue and red ribbons are the corresponding 95% confidence intervals. The dashed black lines are predicted percentiles. Blue circles are individual observations. Several graphs are pointing to a potential model deficiency. Residuals, individuals fits, VPC and npd vs time show an obvious problem with the ability to fit the time course of the data. The model predict a fast and sharp decline from time 0 to 24 hours while the data show a gradual delayed decrease

Figure 5. Core set of diagnostic graphs for Misspecified delay, Effect compartment model. (A) Basic goodness-of-fit plots. Blue points denotes individual data points. Lowess lines are in red. Interpretation should focus on areas where the data are available with little attention to gaps. **(B)** Representative randomly selected individual fits. Blue points denotes individual observed data points, green curve is the IPRED, red curve is the PRED. **(C)** Correlations, distribution and histogram of Empirical Bayes Estimates. **(D)** Simulation-Based Diagnostic Plots (VPC, npd). The blue and red lines are the observed percentiles (10, 50, 90^{th} percentiles), the blue and red ribbons are the corresponding 95% confidence intervals. The dashed

black lines are predicted percentiles. Blue circles are individual observations. Basic goodness-of-fit plots and individual fits provide no hint for model and data disagreement. Simulation-based graphs clearly show that the model can only describe the median and fails to characterize the upper and lower percentiles. The EBE correlation graphs shows high correlations between all PD parameters and this could be an indication of model misspecification.

Figure 6. Core set of diagnostic graphs for Misspecified delay and correlation model (Effect compartment, Full Omega). (A) Basic goodness-of-fit plots. Blue points denotes individual data points. Lowess lines are in red. Interpretation should focus on areas where the data are available with little attention to gaps. (B) Representative randomly selected individual fits. Blue points denotes individual observed data points, green curve is the IPRED, red curve is the PRED. (C) Correlations, distribution and histogram of Empirical Bayes Estimates. (D) Simulation-Based Diagnostic Plots (VPC, npd). The blue and red lines are the observed percentiles (10, 50, 90^{th} percentiles), the blue and red ribbons are the corresponding 95% confidence intervals. The dashed black lines are predicted percentiles. Blue circles are individual observations. The discordance between the 10^{th} and 90^{th} observed percentiles and the corresponding predicted percentiles show that the model may be not be able to characterize the data. High correlation between two parameters may indicates a problem in model parameterization.

Figure 7. Core set of diagnostic graphs for True model – Turn-over model. (A) Basic goodness-of-fit plots for the True Turnover PD model. Blue points denotes individual data points. Lowess lines are in red. Interpretation should focus on areas where the data are available with little attention to the gaps. (B) Representative randomly selected individual fits from the True Turnover PD model. Blue points denotes individual observed data points, green curve is the IPRED, red curve is the PRED. (C) Correlations, distribution histogram and scatterplots of Empirical Bayes Estimates. (D) Simulation-Based Diagnostic Plots (VPC, npd). The blue and red lines are the observed percentiles (10, 50, 90^{th} percentiles), the blue and red ribbons are the corresponding 95% confidence intervals. The dashed black lines are predicted percentiles. Blue circles are individual observations. In general all graphs point to a good model with the exception of

the VPC and npd vs time graphs that are pointing to a potential underestimation of the data 90th percentiles and to a lower extent to an underestimation of the 10th percentiles.

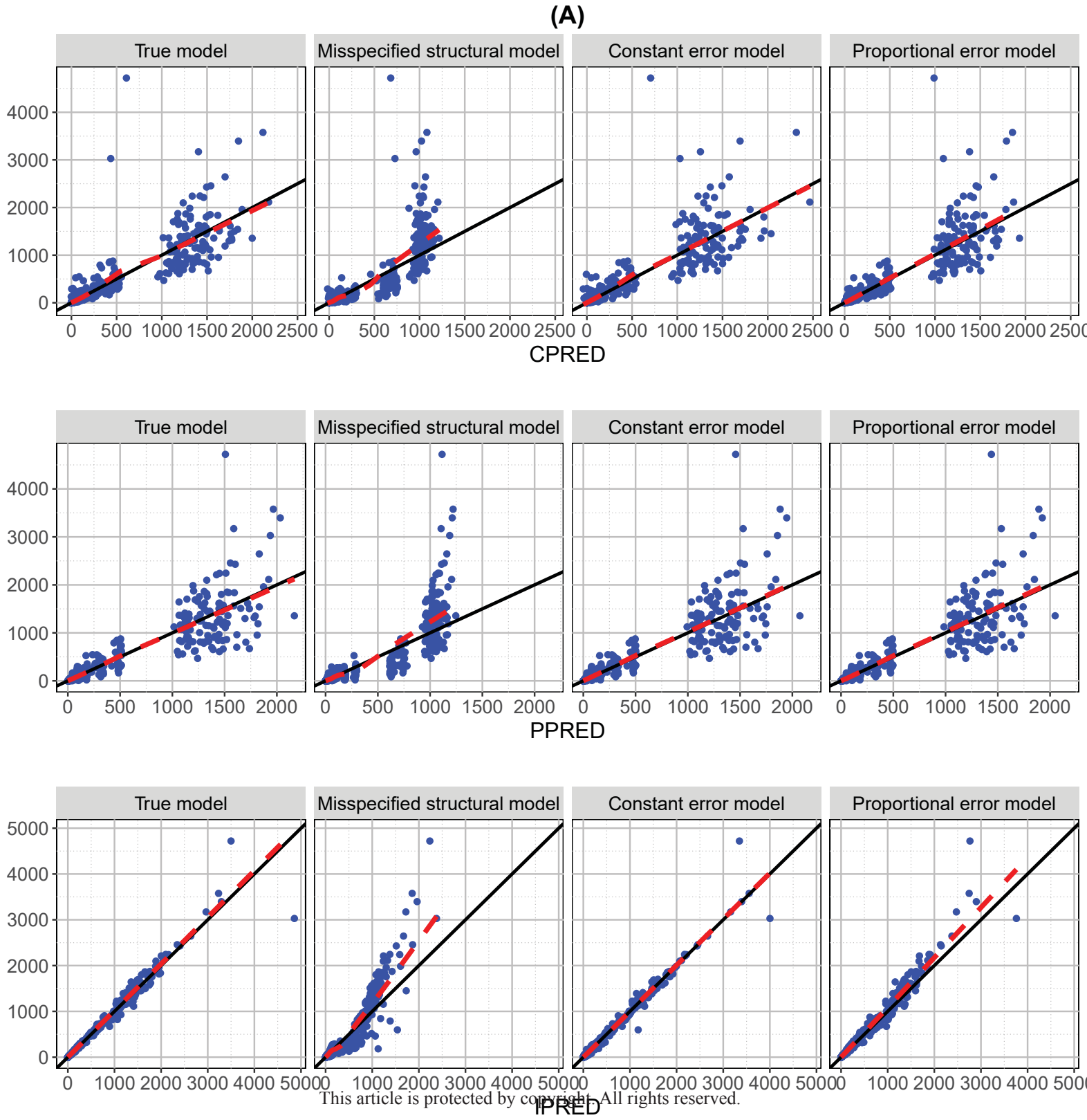
REFERENCES

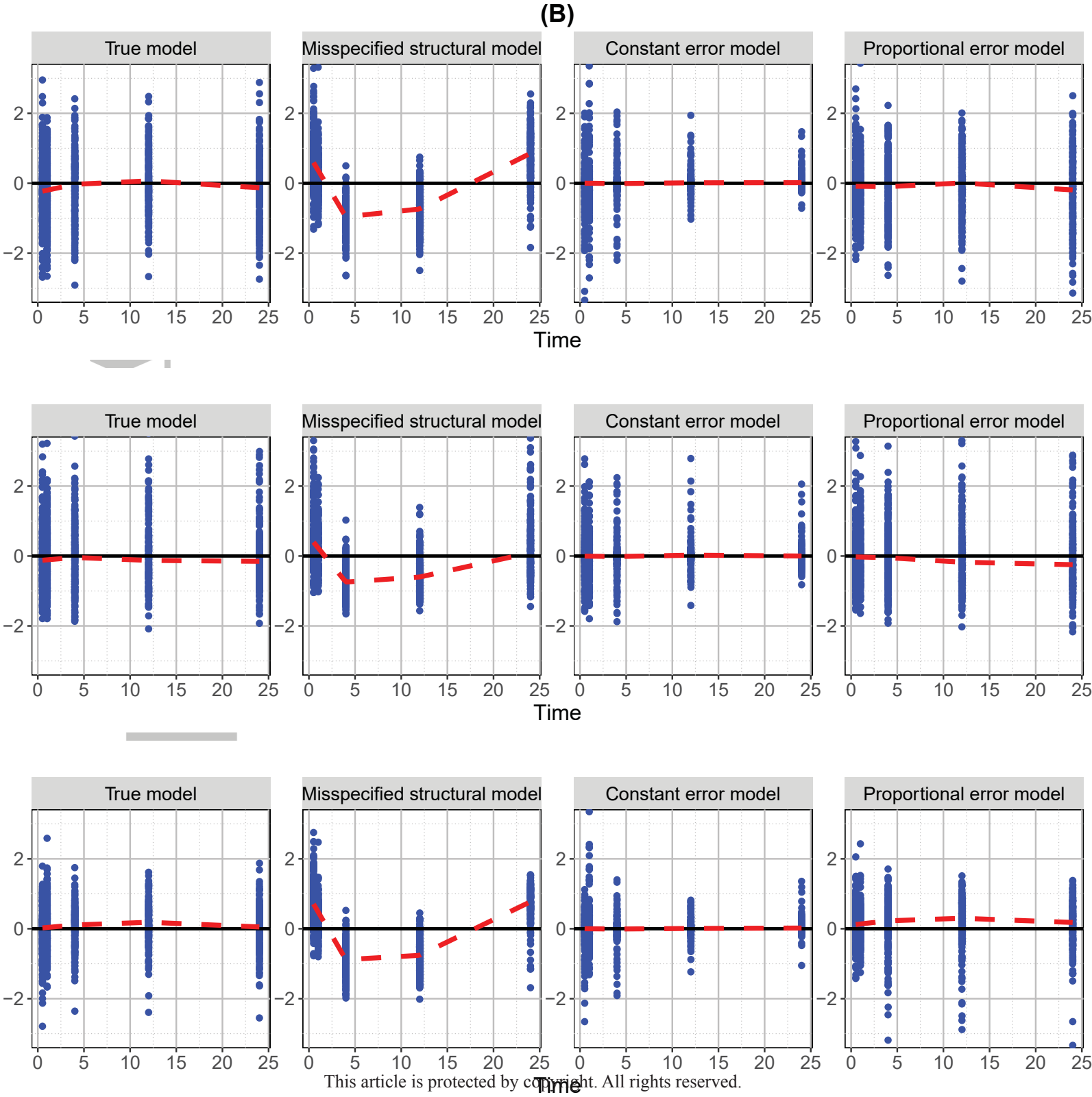
1. Mould, D. R. & Upton, R. N. Basic Concepts in Population Modeling, Simulation, and Model-Based Drug Development. *CPT Pharmacomet. Syst. Pharmacol.* **1**, e6 (2012).
2. Mould, D. R. & Upton, R. N. Basic Concepts in Population Modeling, Simulation, and Model-Based Drug Development—Part 2: Introduction to Pharmacokinetic Modeling Methods. *CPT Pharmacomet. Syst. Pharmacol.* **2**, e38 (2013).
3. Upton, R. N. & Mould, D. R. Basic Concepts in Population Modeling, Simulation, and Model-Based Drug Development: Part 3—Introduction to Pharmacodynamic Modeling Methods. *CPT Pharmacomet. Syst. Pharmacol.* **3**, e88 (2014).
4. Mentré, F. & Escolano, S. Prediction discrepancies for the evaluation of nonlinear mixed-effects models. *J. Pharmacokinet. Pharmacodyn.* **33**, 345–367 (2006).
5. Hooker, A. C., Staatz, C. E. & Karlsson, M. O. Conditional weighted residuals (CWRES): a model diagnostic for the FOCE method. *Pharm. Res.* **24**, 2187–2197 (2007).
6. 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.* **23**, 2036–2049 (2006).
7. Karlsson, M. O. & Savic, R. M. Diagnosing Model Diagnostics. *Clin. Pharmacol. Ther.* **82**, 17–20 (2007).
8. Holford, N. VPC: the visual predictive check superiority to standard diagnostic (Rorschach) plots. **PAGE 14**, Abstr 738 [www.page-meeting.org/?abstract=738] (2005).
9. Ma, G., Olsson, B. L., Rosenborg, J. & Karlsson, M. O. Quantifying Lung Function Progression in Asthma. **PAGE 18**, Abstr 1562 [www.page-meeting.org/?abstract=1562] (2009).

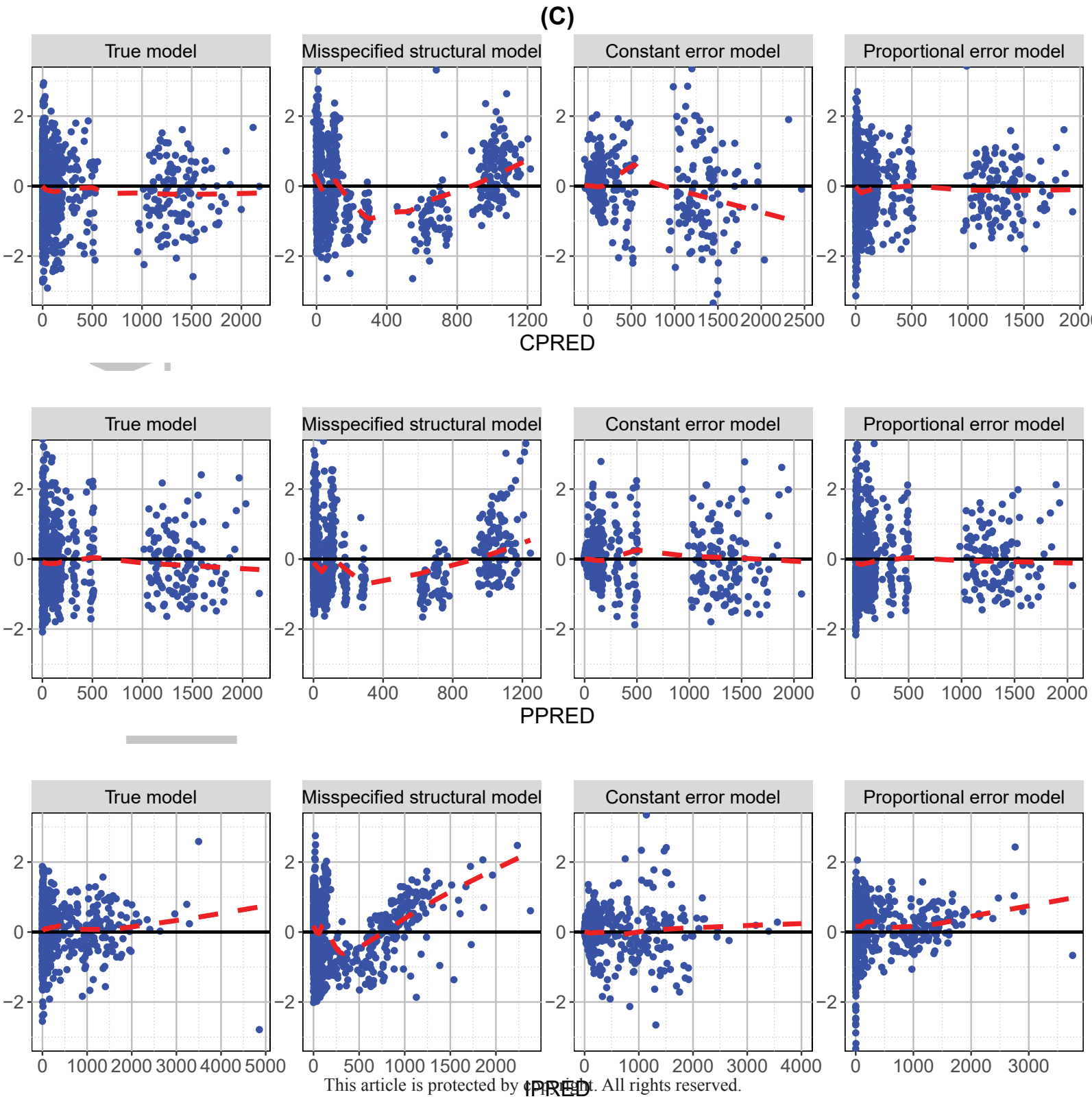
10. Savic, R. M. & Karlsson, M. O. Importance of shrinkage in empirical bayes estimates for diagnostics: problems and solutions. *AAPS J.* **11**, 558–569 (2009).
11. Yano, Y., Beal, S. L. & Sheiner, L. B. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. *J. Pharmacokinet. Pharmacodyn.* **28**, 171–192 (2001).
12. Bonate, P. L. *Pharmacokinetic-pharmacodynamic modeling and simulation*. (Springer, 2006).
13. Acharya, C., Hooker, A. C., Jönsson, S. & Karlsson, M. O. A diagnostic tool for population models using non-compartmental analysis: nca_ppc functionality for R. **PAGE 23**, Abstr 3103 [www.page-meeting.org/?abstract=3103] (2014).
14. Largajolli, A., Jönsson, S. & Karlsson, M. O. The OFVPPC: A simulation objective function based diagnostic. **PAGE 23**, Abstr 3208 [www.page-meeting.org/?abstract=3208] (2014).
15. Jadhav, P. R. & Gobburu, J. V. S. A new equivalence based metric for predictive check to qualify mixed-effects models. *AAPS J.* **7**, E523–E531 (2005).
16. Laffont, C. M. & Concordet, D. A New Exact Test for the Evaluation of Population Pharmacokinetic and/or Pharmacodynamic Models Using Random Projections. *Pharm. Res.* **28**, 1948–1962 (2011).
17. Comets, E. & Brendel, K. Model evaluation in nonlinear mixed effect models, with applications to pharmacokinetics. *J. Société Fr. Stat.* **151**, 106–128 (2010).
18. Lavielle, M. & Bleakley, K. Automatic data binning for improved visual diagnosis of pharmacometric models. *J. Pharmacokinet. Pharmacodyn.* **38**, 861–871 (2011).
19. Sonehag, C., Olofsson, N., Simander, R., Nordgren, R. & Harling, K. Automatic binning for visual predictive checks. **PAGE 23**, Abstr 3085 [www.page-meeting.org/?abstract=3085] (2014).
20. Bergstrand, M., Hooker, A. C., Wallin, J. E. & Karlsson, M. O. Prediction-Corrected Visual Predictive Checks for Diagnosing Nonlinear Mixed-Effects Models. *AAPS J.* **13**, 143–151 (2011).

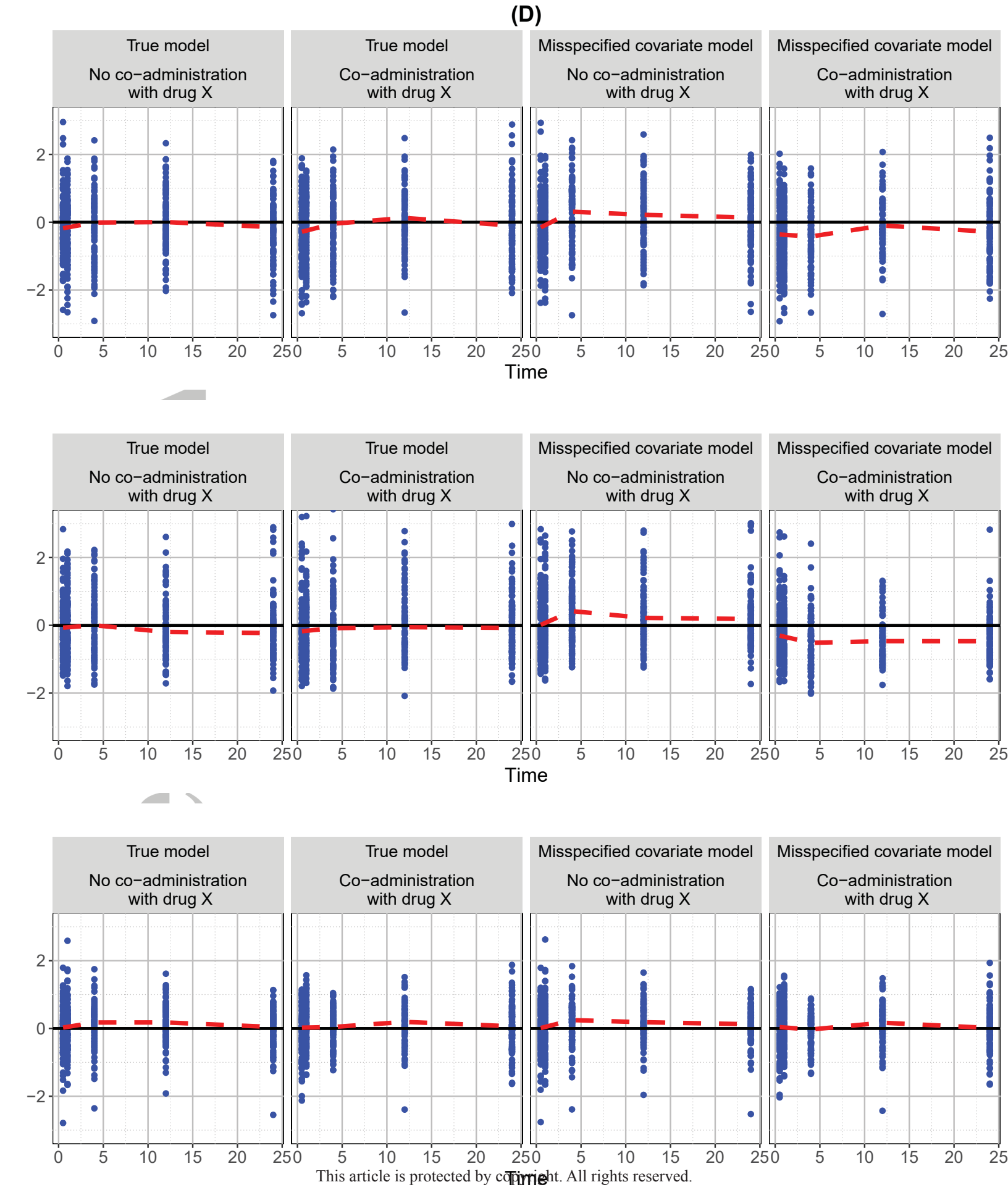
21. McCune, J. S. *et al.* Busulfan in Infant to Adult Hematopoietic Cell Transplant Recipients: A Population Pharmacokinetic Model for Initial and Bayesian Dose Personalization. *Clin. Cancer Res.* **20**, 754–763 (2014).
22. Wilkins, J., Karlsson, M. & Jonsson, E. Patterns and power for the visual predictive check. 15-2006. **PAGE 15**, Abstr 1029 [www.page-meeting.org/?abstract=1029] (2006).
23. Comets, E., Nguyen, T. H. T. & Mentré, F. Additional features and graphs in the new npde library for R. **PAGE 22**, Abstr 2775 [http://www.page-meeting.org/default.asp?abstract=2775] (2013).
24. O'Reilly, R. A., Aggeler, P. M. & Leong, L. S. Studies on the coumarin anticoagulant drugs: the pharmacodynamics of warfarin in man. *J. Clin. Invest.* **42**, 1542–1551 (1963).
25. O'Reilly, R. A. & Aggeler, P. M. Studies on Coumarin Anticoagulant Drugs. Initiation of warfarin therapy without a loading dose. *Circulation* **38**, 169–177 (1977).
26. Leary, R., Dunlavey, M., Chittenden, J., Matzuka, B. & Guzy, S. QRPEM—a new standard of accuracy, precision, and efficiency in NLME population PK/PD methods. *Pharsight® Certara™ Co.* (2011).
27. Hengl, S., Kreutz, C., Timmer, J. & Maiwald, T. Data-based identifiability analysis of non-linear dynamical models. *Bioinforma. Oxf. Engl.* **23**, 2612–2618 (2007).
28. Thai, H.-T., Mentré, F., Holford, N. H. G., Veyrat-Follet, C. & Comets, E. Evaluation of bootstrap methods for estimating uncertainty of parameters in nonlinear mixed-effects models: a simulation study in population pharmacokinetics. *J. Pharmacokinet. Pharmacodyn.* **41**, 15–33 (2014).
29. Chen, J. & Chen, Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771 (2008).
30. Delattre, M., Lavielle, M. & Poursat, M.-A. A note on BIC in mixed-effects models. *Electron. J. Stat.* **8**, 456–475 (2014).

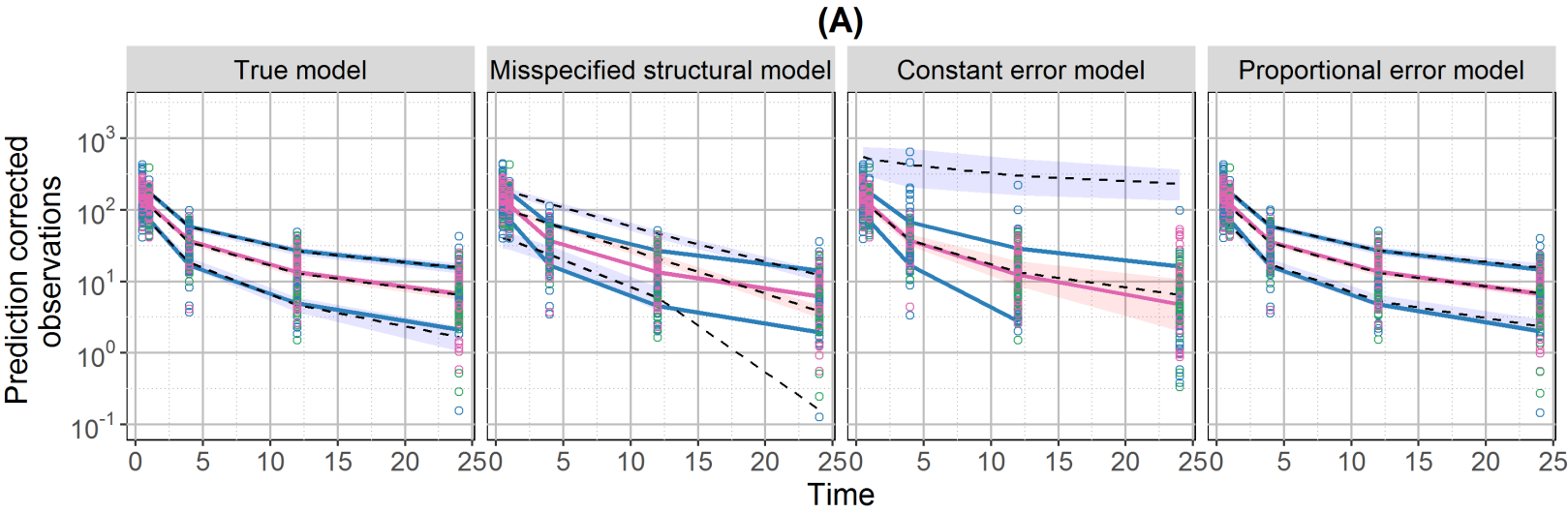
31. Mould, D. R. & Frame, B. Population Pharmacokinetic-Pharmacodynamic Modeling of Biological Agents: When Modeling Meets Reality. *J. Clin. Pharmacol.* **50**, 91S–100S (2010).
32. Friberg, L. E., Greef, R. de, Kerbusch, T. & Karlsson, M. O. Modeling and simulation of the time course of asenapine exposure response and dropout patterns in acute schizophrenia. *Clin. Pharmacol. Ther.* **86**, 84–91 (2009).
33. Björnsson, M. A. & Simonsson, U. S. H. Modelling of pain intensity and informative dropout in a dental pain model after naproxcinod, naproxen and placebo administration. *Br. J. Clin. Pharmacol.* **71**, 899–906 (2011).
34. Comets, E., Brendel, K. & Mentré, F. Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: the npde add-on package for R. *Comput. Methods Programs Biomed.* **90**, 154–166 (2008).
35. Samtani, M. N., Perez-Ruixo, J. J., Brown, K. H., Cerneus, D. & Molloy, C. J. Pharmacokinetic and pharmacodynamic modeling of pegylated thrombopoietin mimetic peptide (PEG-TPOm) after single intravenous dose administration in healthy subjects. *J. Clin. Pharmacol.* **49**, 336–350 (2009).
36. Nguyen, T. H. T., 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.* **39**, 499–518 (2012).

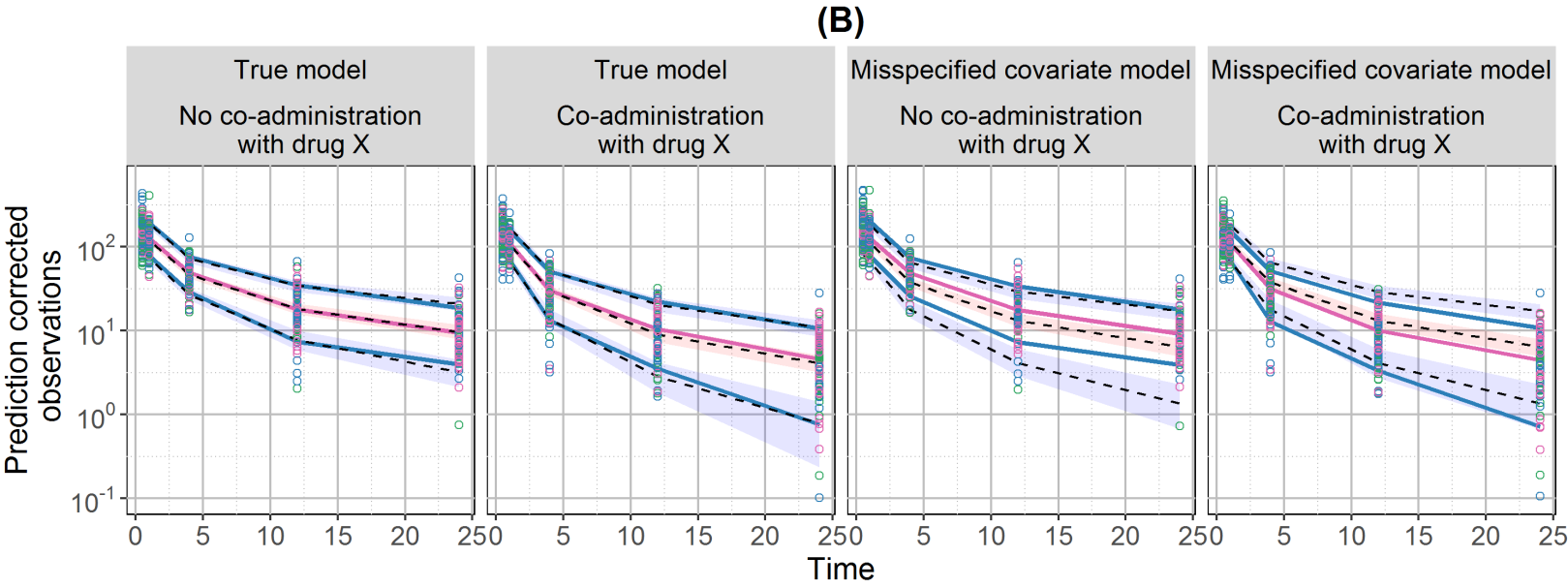


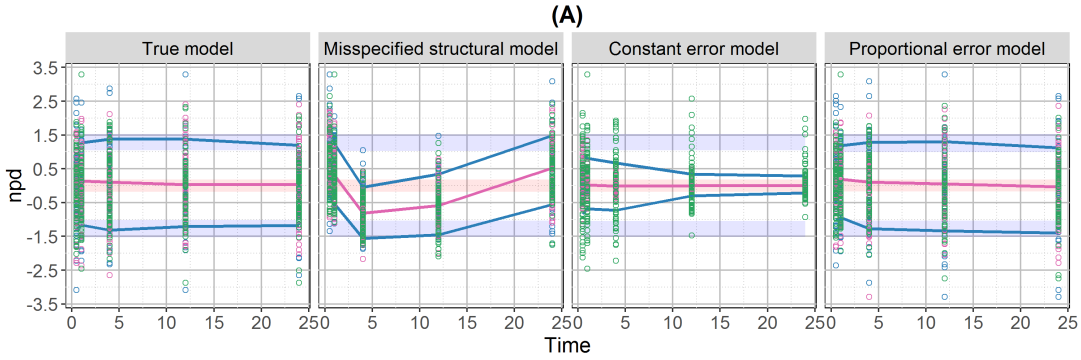


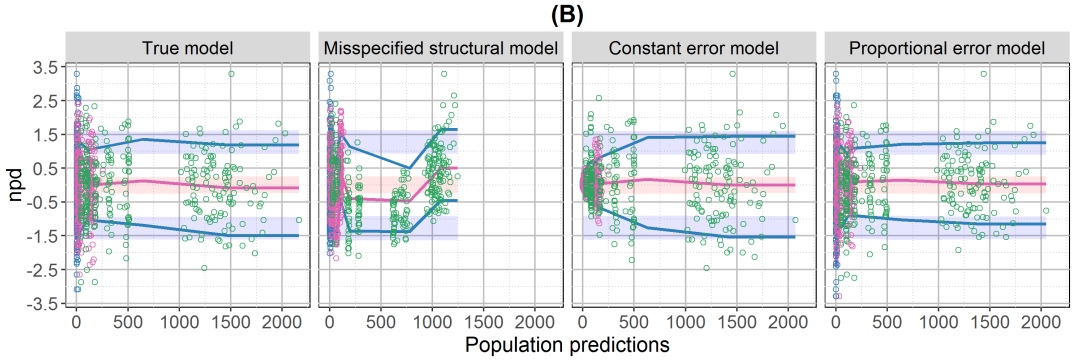


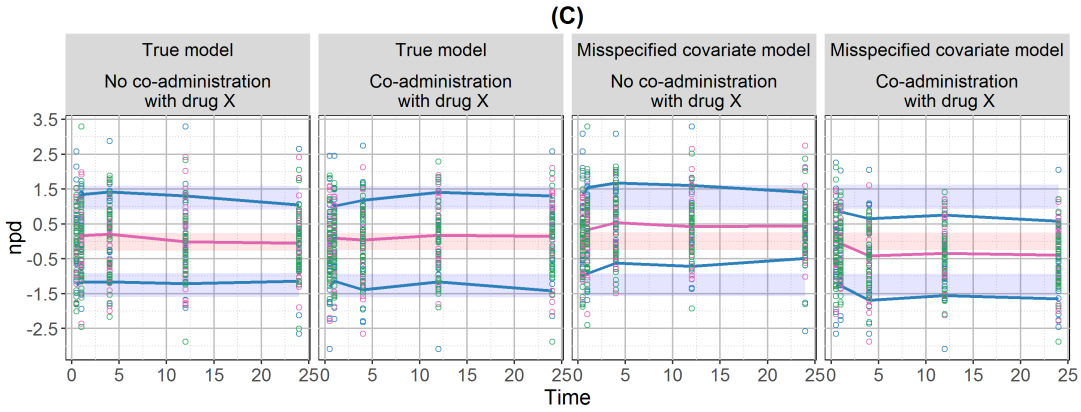


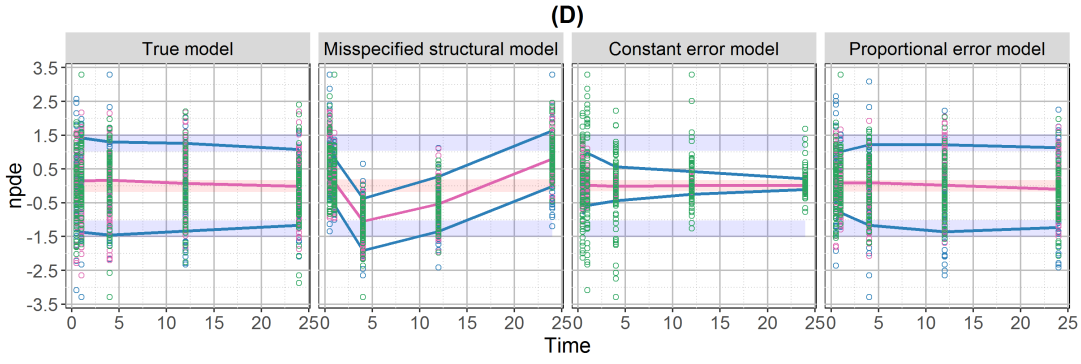


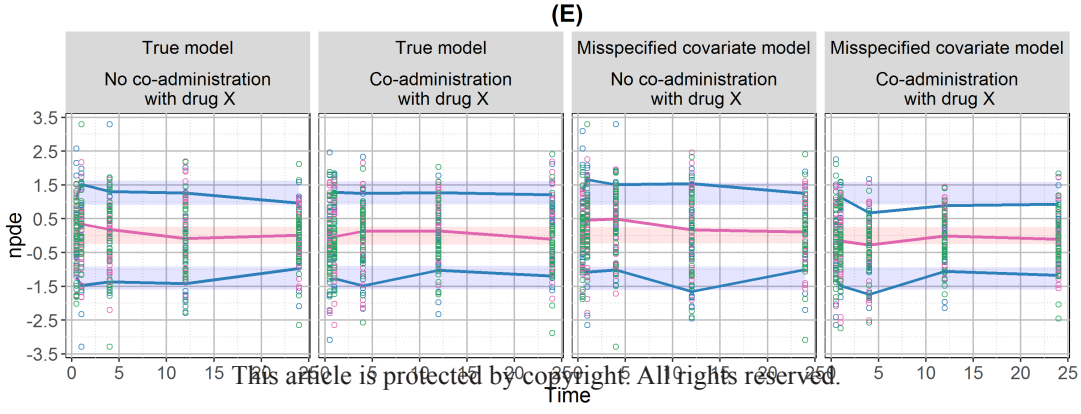




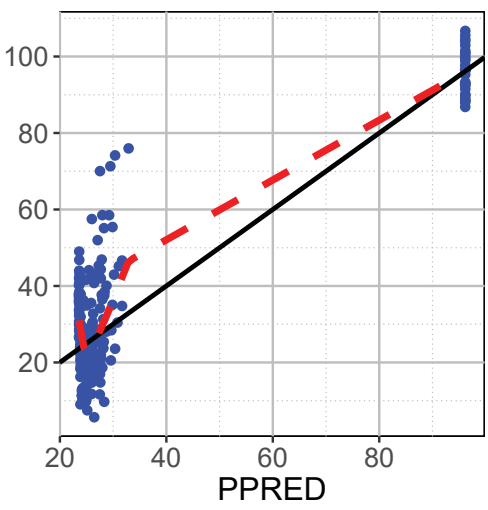


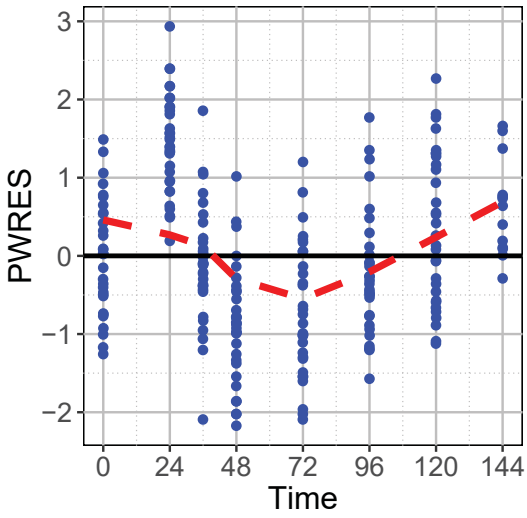


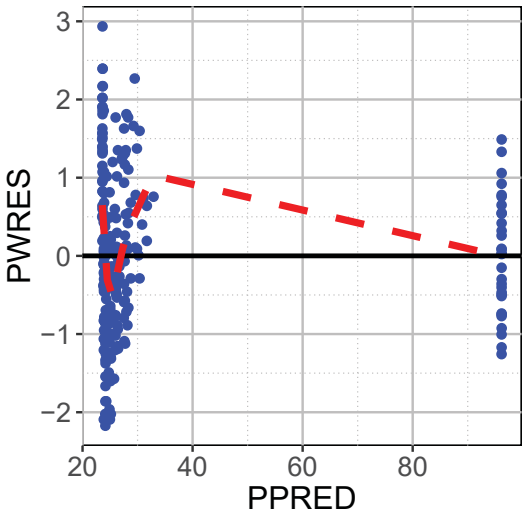


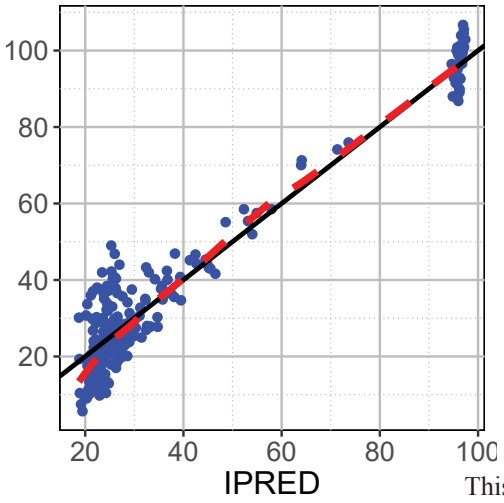


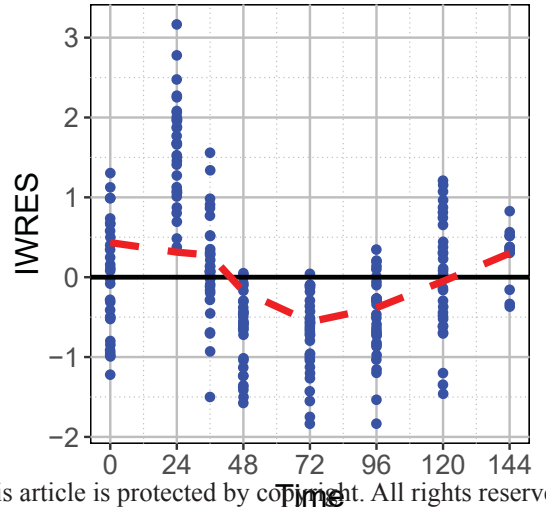
(A) Basic Goodness−of−fit Plots (Misspecified delay, Immediate effect)

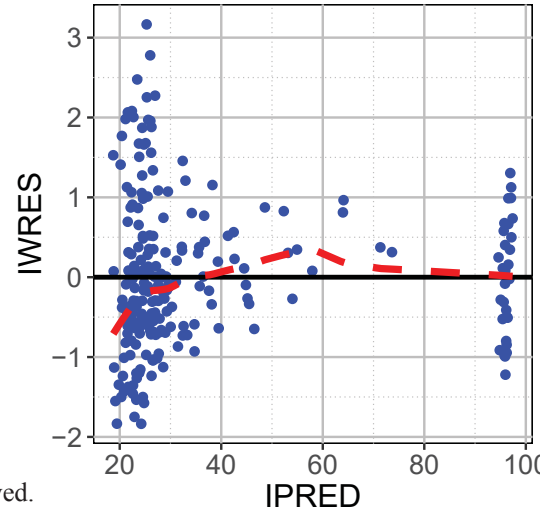




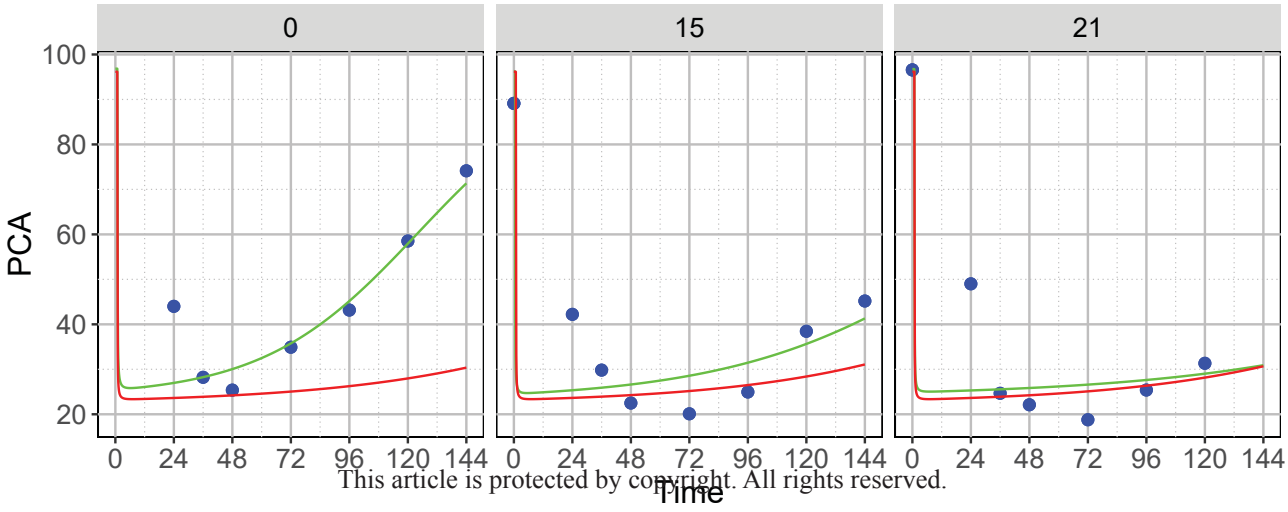




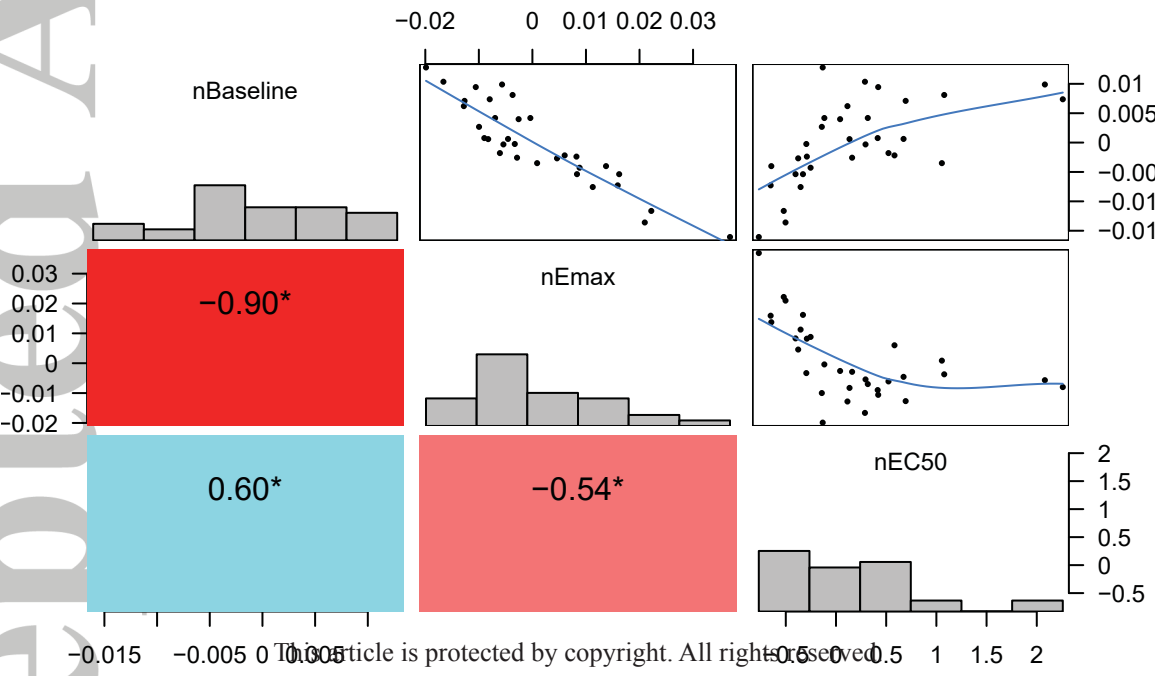




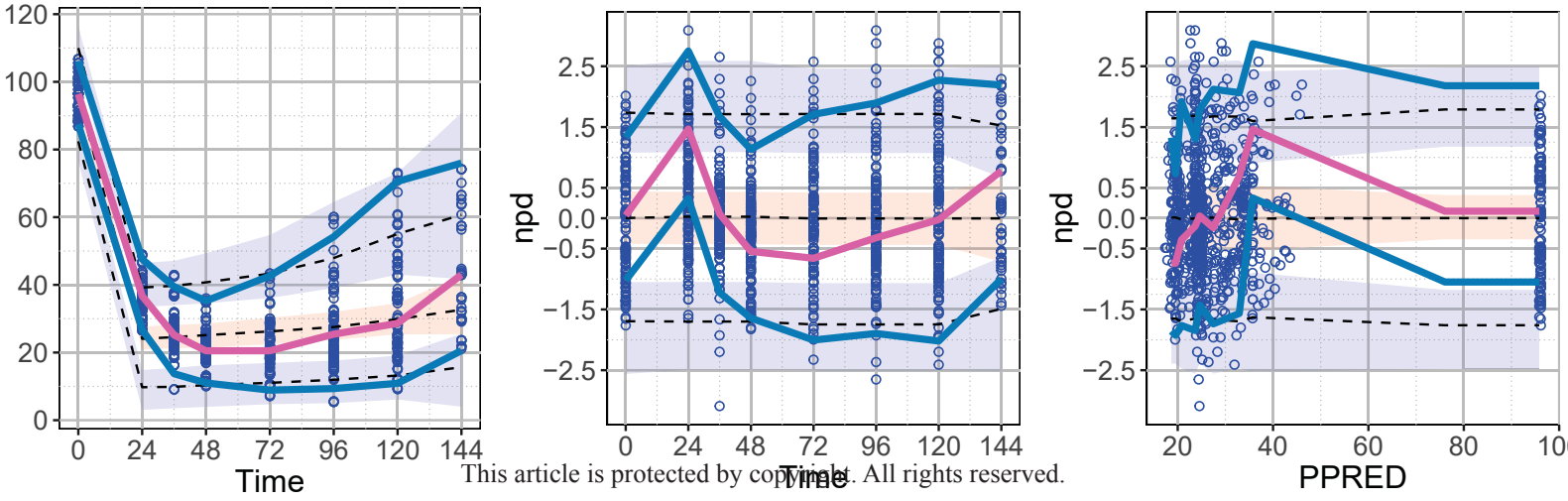
(B) Representative Individual Fits (Misspecified delay, Immediate effect)



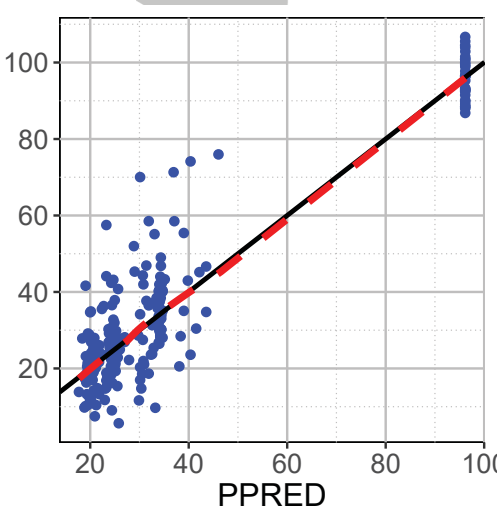
(C) Correlations, histogram of EBE (Misspecified delay, Immediate effect)

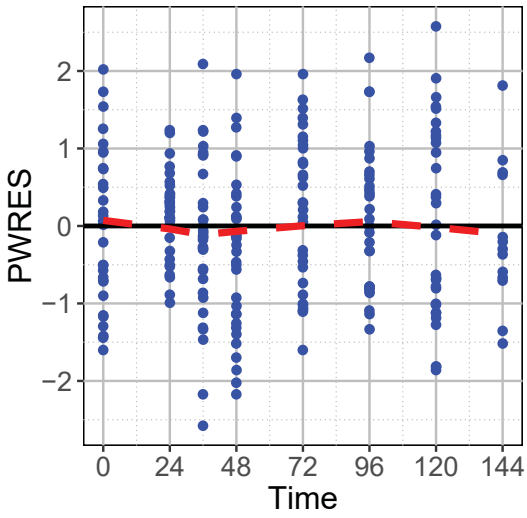


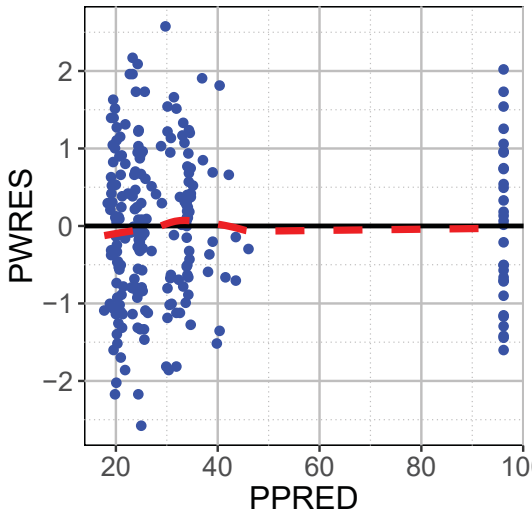
(D) Simulation−based Goodness−of−fit Plots (Misspecified delay, Immediate effect)

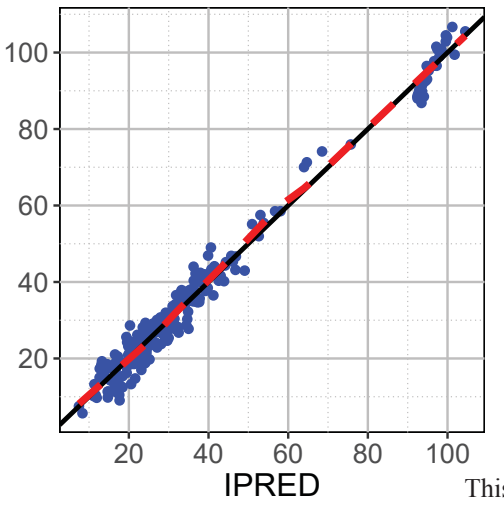


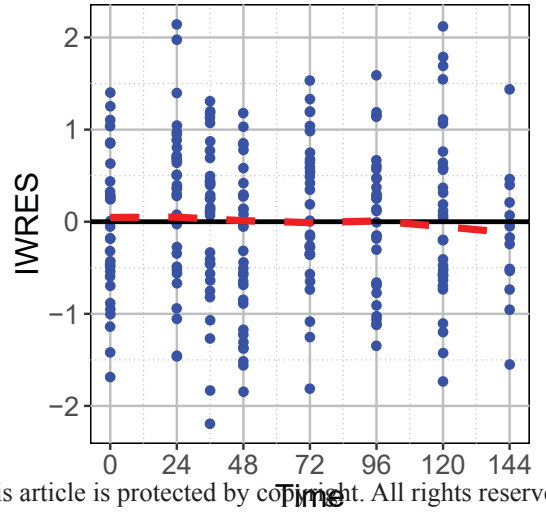
(A) Basic Goodness−of−fit Plots (Misspecified delay, Effect compartment)

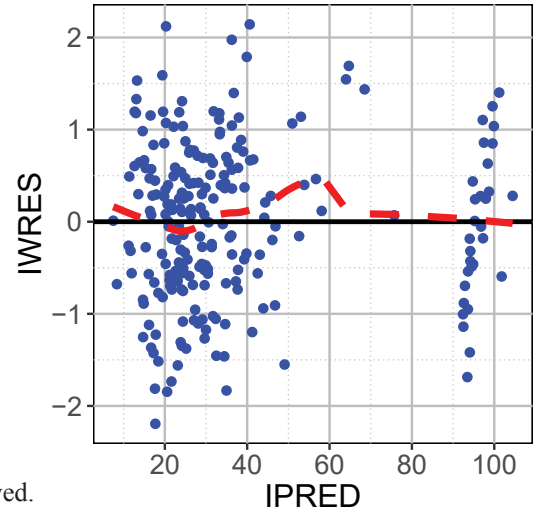




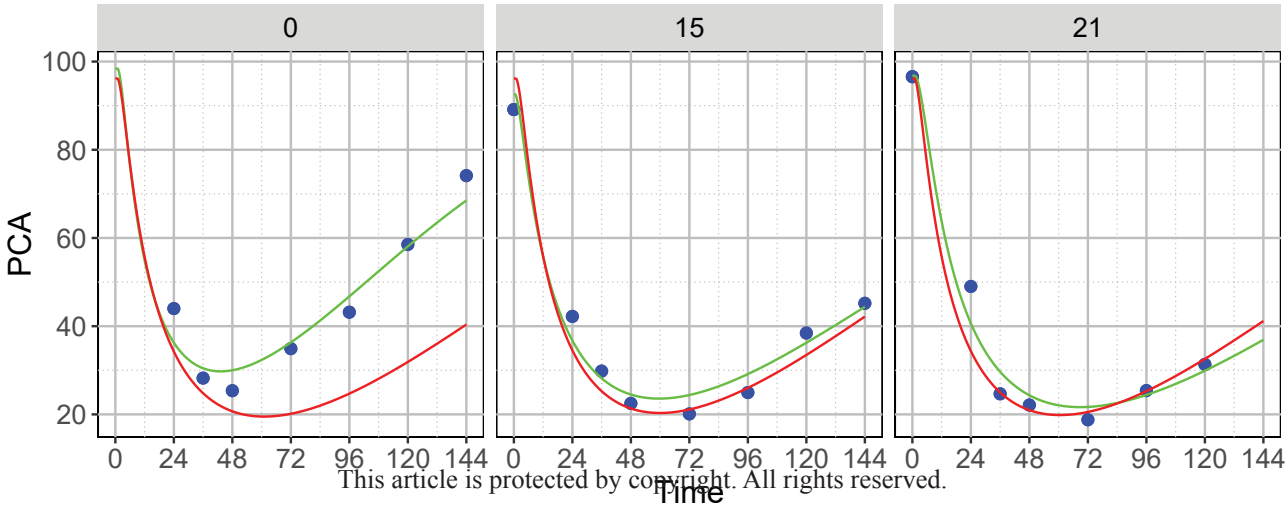




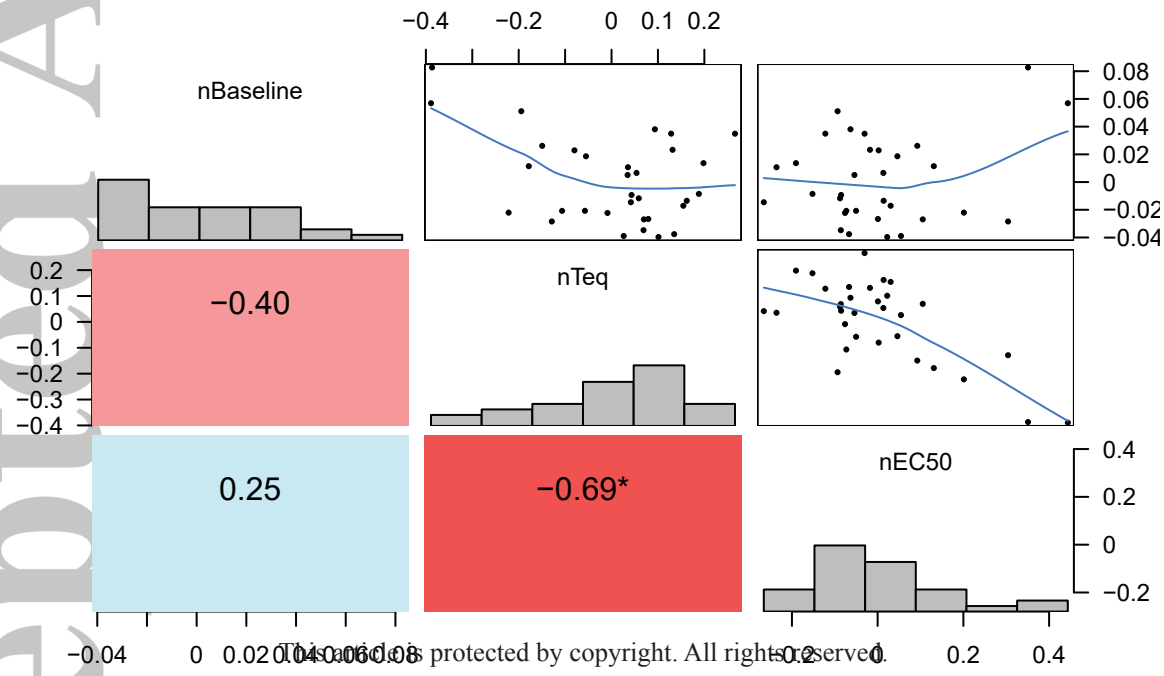




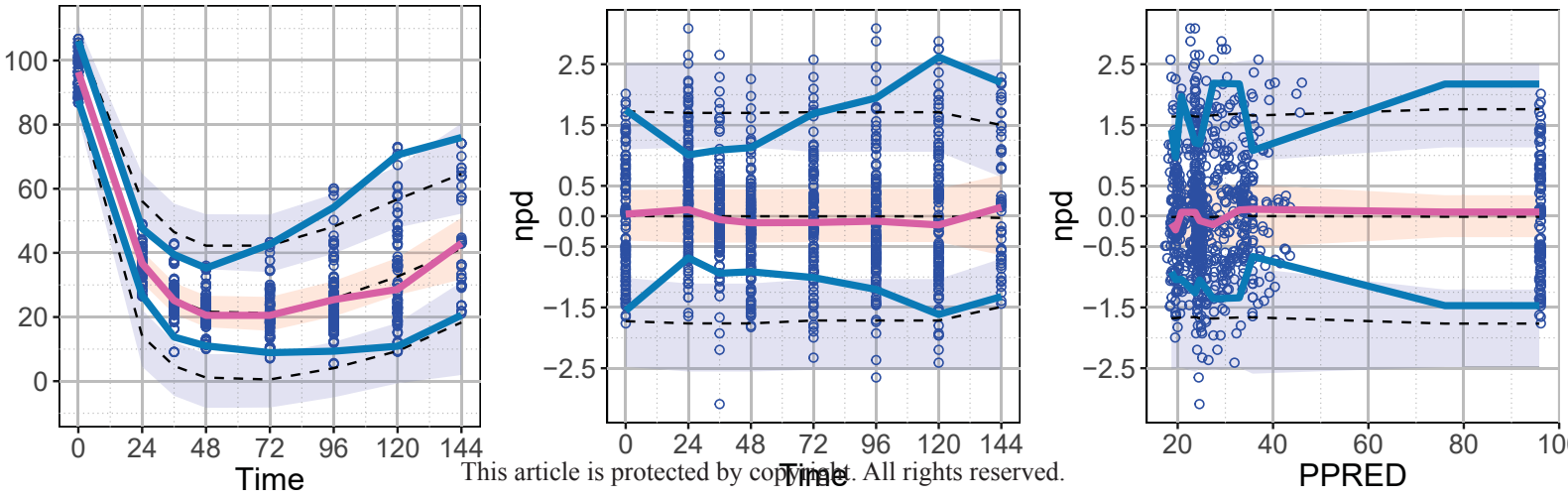
(B) Representative Individual Fits (Misspecified delay, Effect compartment)



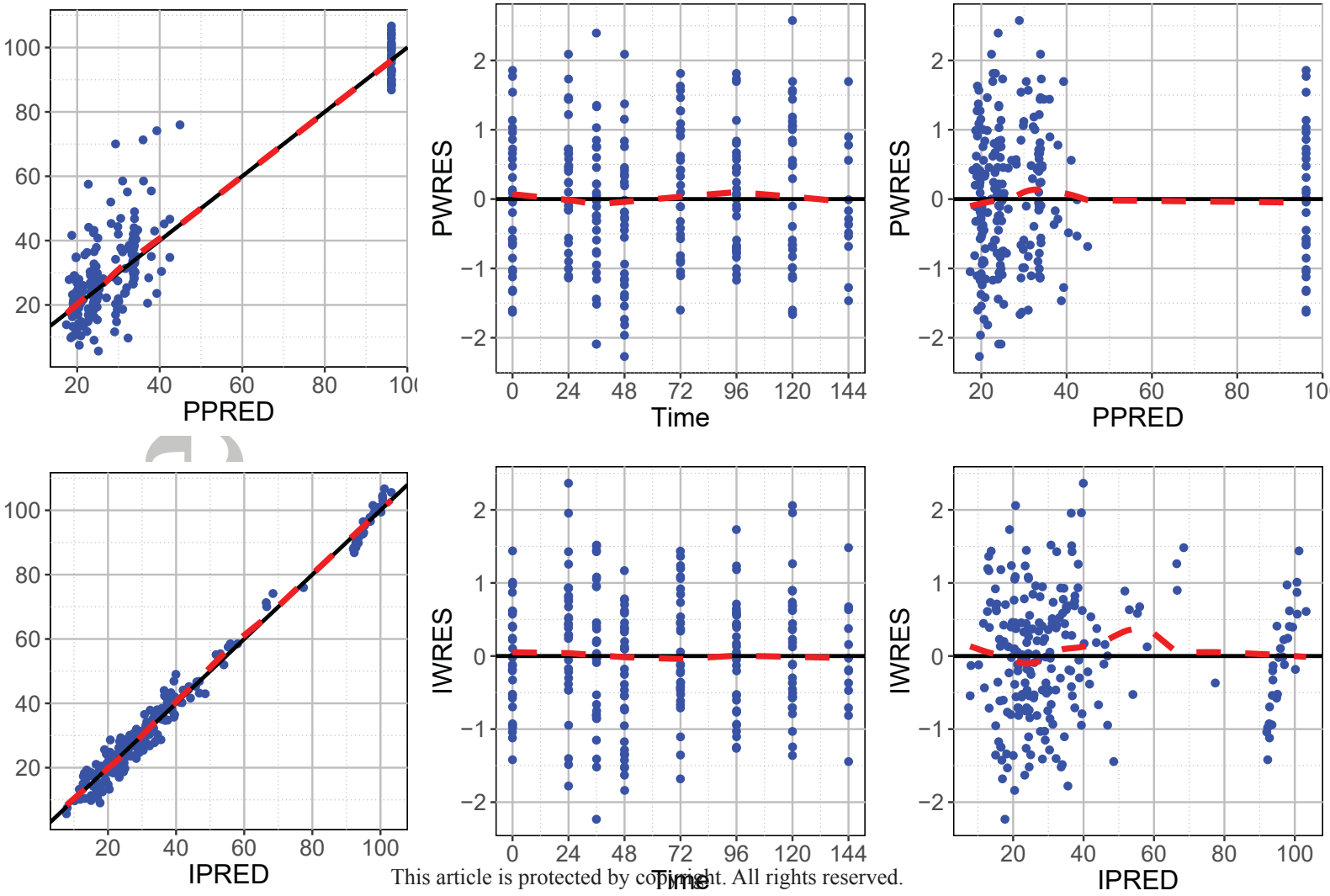
(C) Correlations, histogram of EBE (Misspecified delay, Effect compartment)



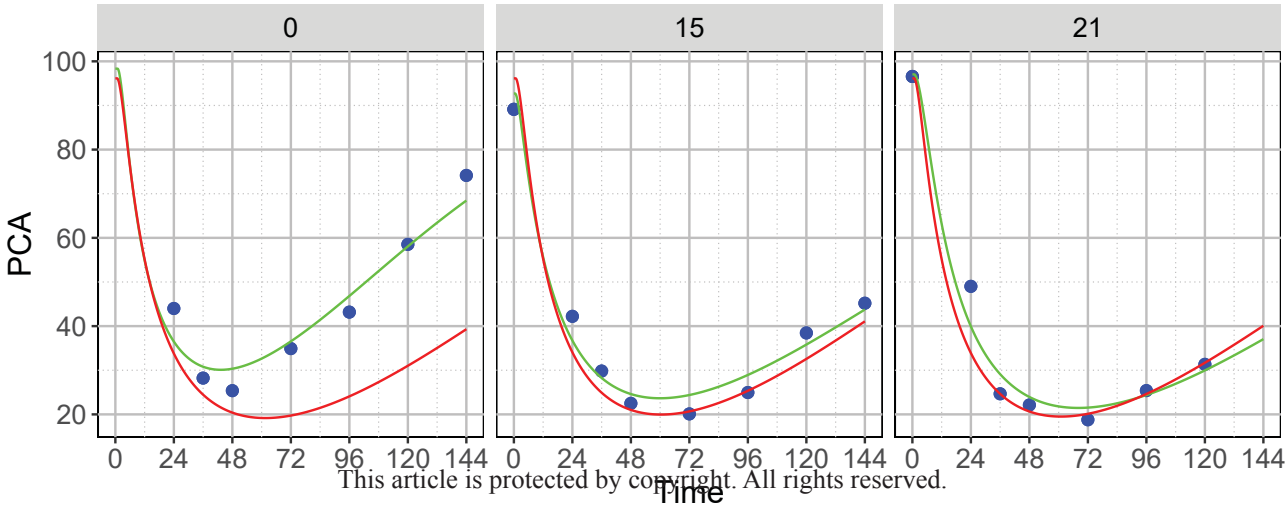
(D) Simulation−based Goodness−of−fit Plots (Misspecified delay, Effect compartment)



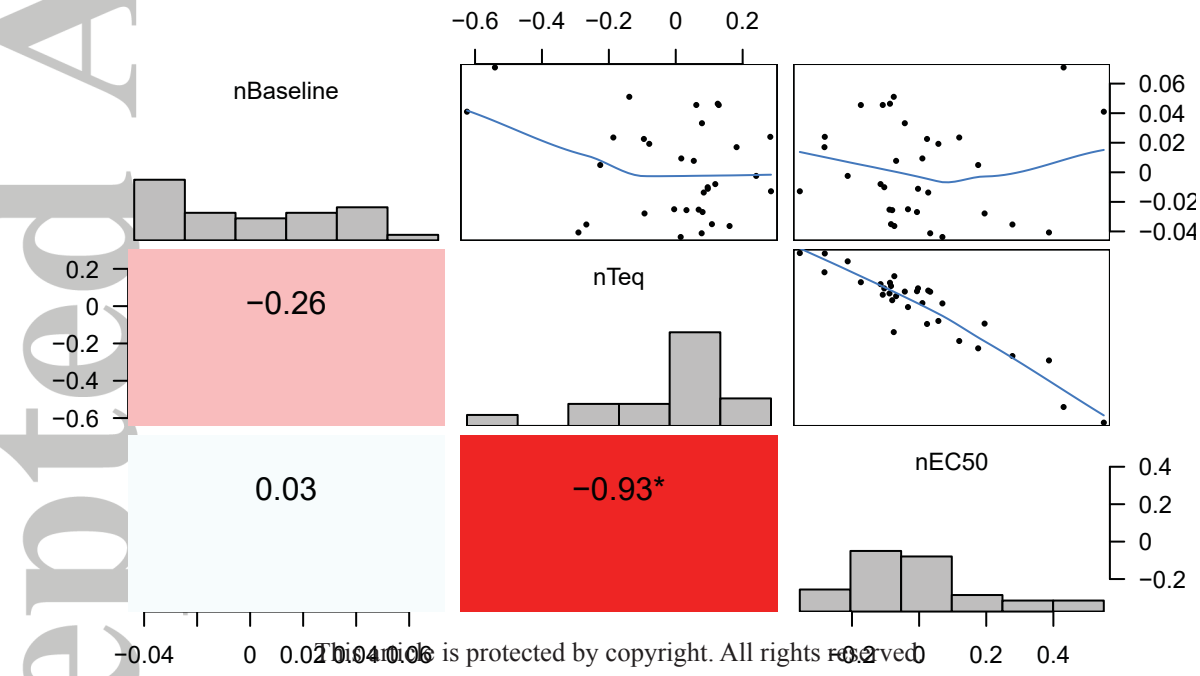
(A) Basic Goodness−of−fit Plots (Misspecified delay and correlation)



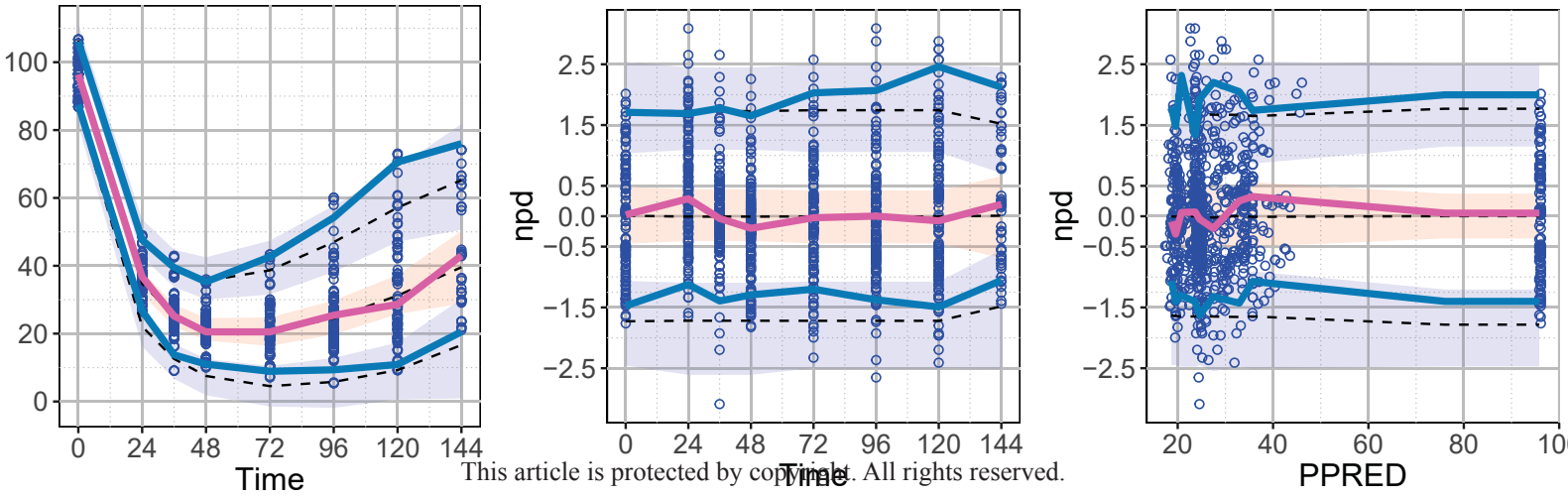
(B) Representative Individual Fits (Misspecified delay and correlation)

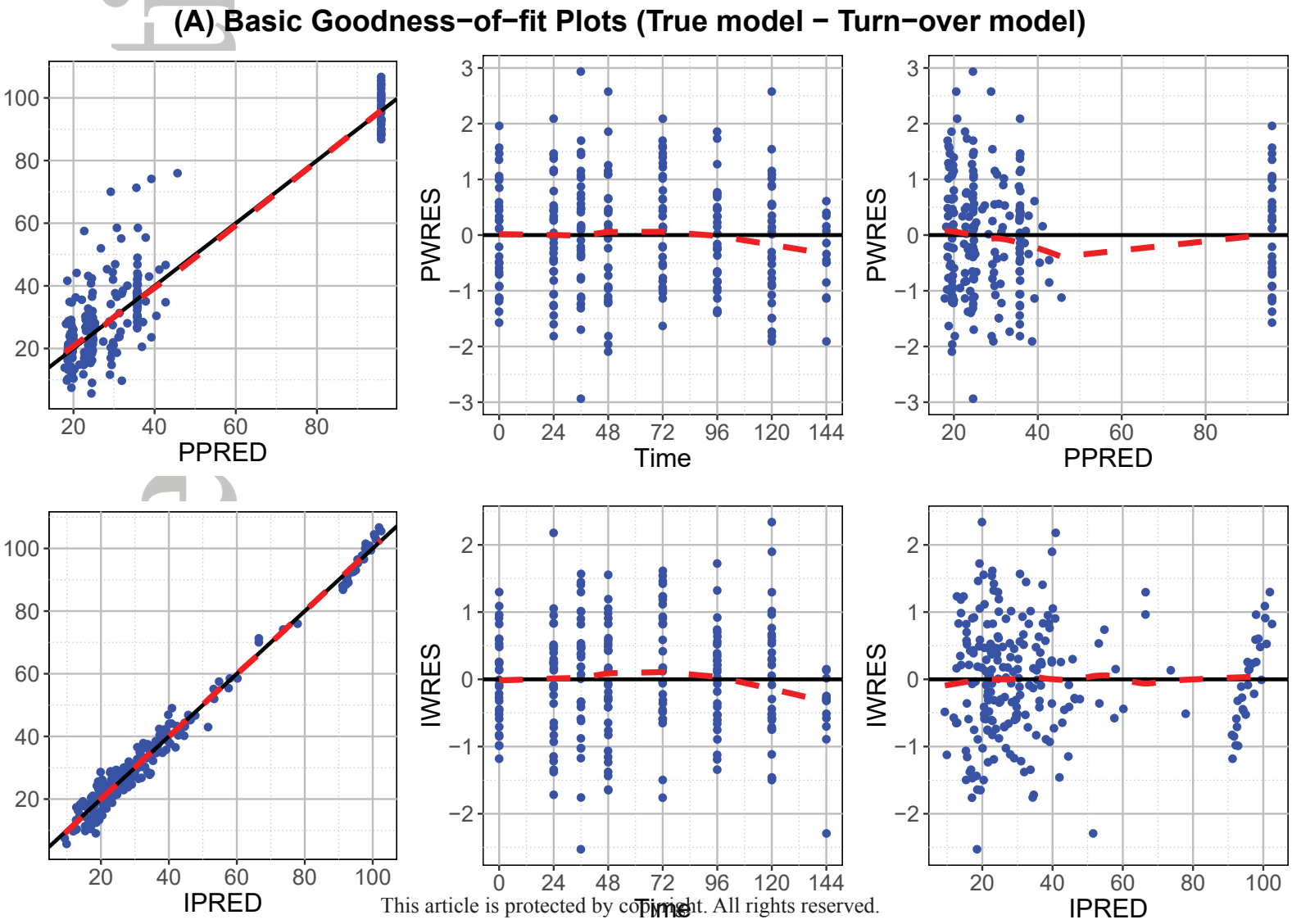


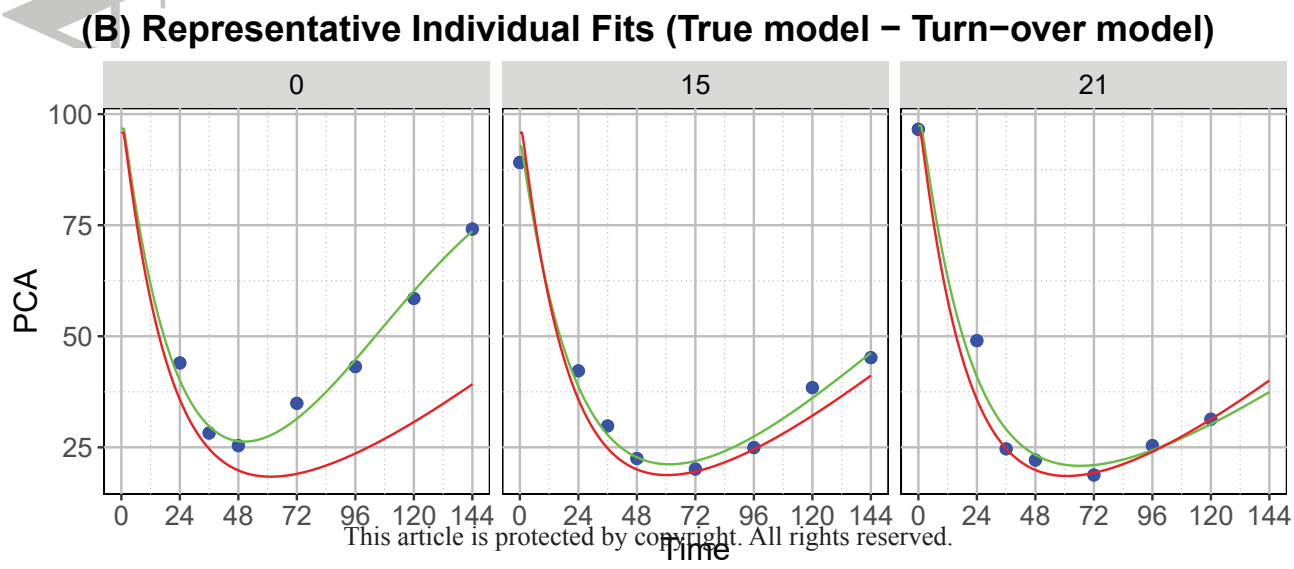
(C) Correlations, histogram of EBE (Misspecified delay and correlation)

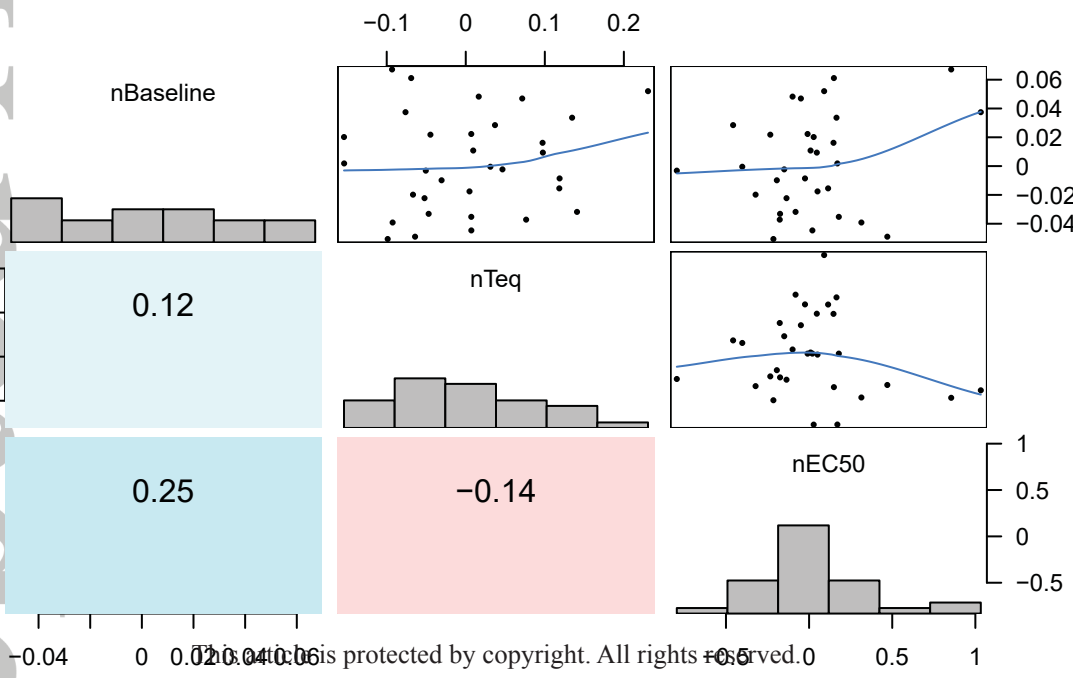


(D) Simulation−based Goodness−of−fit Plots (Misspecified delay and correlation)









(D) Simulation−based Goodness−of−fit Plots (True model − Turn−over model)

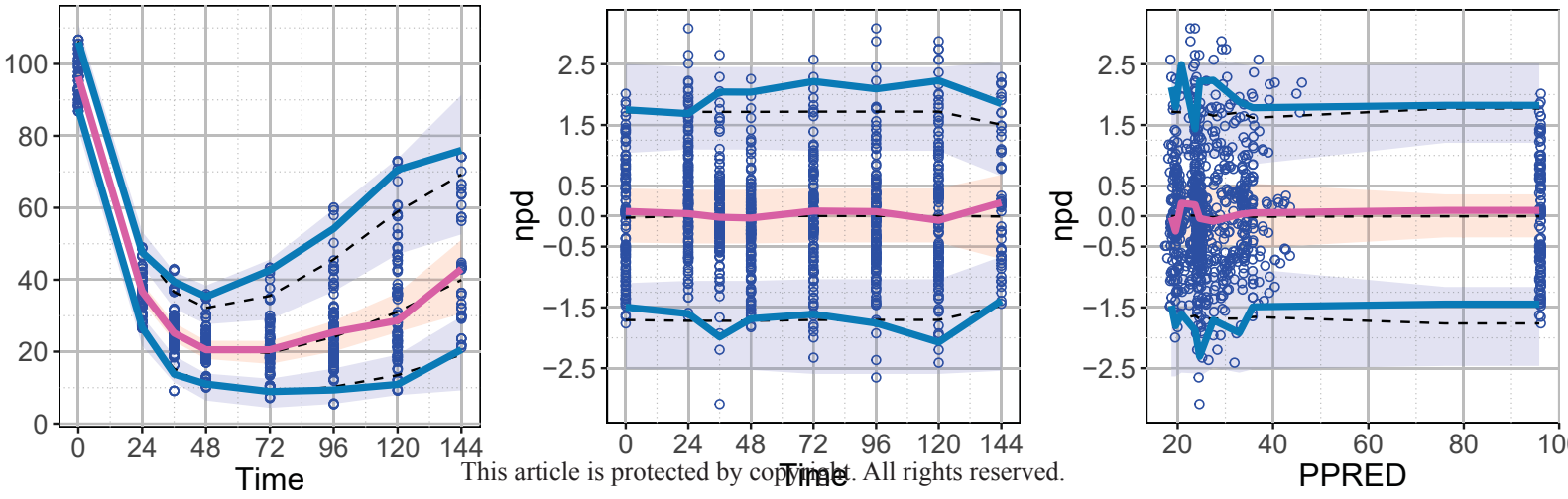


Table 1. Various evaluation graphs in NLMEM[1] and proposal for a core set of evaluation graphs

GRAPHS	In core set	What to expect if the model is correct?	What to do if the graph does not fulfill the requirements?
BASIC (PREDICTION-BASED) EVALUATION			
POPULATION-BASED GRAPHS			
OBS vs xPRED (x= ∅, C, $P^{(2)}$)	YES NB: CPRED or PPRED[3]	Data points are scattered around the identity line (but not necessarily evenly).	Trends may suggest a modification of structural model, residual error model or inter-individual variability model. NB: Trends can also appear in absence of model misspecifications for models highly nonlinear with respect to random effects and large inter-individual variability, especially when using $PRED_{FO}$.
xWRES (x= ∅, C, $P^{(2)}$) vs time or xPRED	YES NB: CWRES or PWRES[3]	Data points are scattered around the horizontal zero-line (more or less evenly).	Trends may suggest a modification of structural model, residual error model or inter-individual variability model. Trends by conditioning on covariates suggest including covariates. NB: Trends can also appear in absence of model misspecifications for models highly nonlinear with respect to random effects and large inter-individual variability, especially when using WRES.
xWRES (x= ∅, C, $P^{(2)}$*) vs covariates	YES if covariates are considered	No substantial correlation appears	Trends suggest including covariates or changing the covariate model
INDIVIDUAL-BASED GRAPHS[4]			
Individual fits	YES NB: at least for some representative individuals	Observations are distributed evenly around the individual predicted curve.	A substantial discordance between observations and predictions suggests a modification of the structural model, parameter variability model or the residual error model. NB: This diagnostic is not useful for sparse data.
OBS vs IPRED	YES	Data points are scattered	Trends may suggest a modification of the structural model or the

		evenly around the identity line. Points cluster closer to the line than with Observations vs PRED, especially when inter-individual variability is large	residual error model. A lack of trend may not be necessarily be associated with absence of model misspecification if data are sparse.
IWRES vs time or IPRED	YES NB: Graphs of absolute IWRES vs IPRED are also informative	Data points are scattered evenly around the horizontal zero-line. Most of the points lie within (-1.96;1.96)	Trends suggest a modification of structural model or residual error model. A cone-shaped graphs of IWRES vs IPRED suggests a change in the error model. A lack of trend may not be necessarily be associated with absence of model misspecification if data are sparse.
Correlation between EBEs	YES	No trend is expected in model without correlation between random effects if data are rich (i.e. low eta-shrinkage)	Correlation between EBE suggests including correlation between random effects unless data are sparse.
EBEs vs covariates	YES if covariates are considered	No substantial correlation appears between EBE and covariates	Trends between EBE and covariates suggest including covariates or changing the covariate model
SIMULATION-BASED GRAPHS			
VPC or pcVPC	YES NB: The choice between the two depends on the importance of covariates and use of adaptive designs	Observed percentiles are not systematically different from the corresponding predicted percentiles and are within the corresponding confidence interval	Trends may suggest a modification of the structural model, the residual error model or the parameter variability model. Trends when conditioning on covariates suggest including covariates or changing the covariate model.
NPC coverage		Observed percentiles are within the confidence interval of the corresponding	Trends may suggest a modification of the structural model, the residual error model or the inter-individual variability model. Trends when conditioning on covariates suggest including

		predicted percentiles	covariates.
npd (npde) vs time or PPRED[5]	YES	Data points are scattered evenly around the horizontal zero-line. Most of the points lie within (-1.96;1.96). For graphs with observed and predicted percentiles and the confidence intervals: Observed percentiles are not systematically different from the corresponding predicted percentiles and are within the corresponding confidence interval	Trends may suggest a modification of structural model, residual error model or inter-individual variability models. A cone-shaped graph if npd vs PPRED suggest a change in the residual error model Trends when conditioning on covariates suggest including covariates.
npd (npde) vs covariates	YES if covariates are considered	No substantial correlation appears	Trends suggest including covariates or changing the covariate model

[1]See text or **Supplementary table S3** for definition of terms
[2]PPRED and PWRES are denoted EPRED and EWRES in NONMEM, respectively[3] Depending on the method used for parameter estimation (see text)
[4] Caution in interpretation in case of high shrinkage
[5]npd is preferred over npde for graphical use