Model evaluation in nonlinear mixed effect models

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

[1] *INSERM UMR738, Paris, France; Université Paris Diderot, Paris, France* [2] *IRIS, Servier, Courbevoie, France*

Objective: Model evaluation is an important part of model building, and has been the subject of regulatory guidelines. We illustrate the use of some recently proposed metrics on several simulated datasets.

Introduction

- Several simulation-based metrics developed over the last decade:
  - Visual Predictive Checks (VPC) [1]
  - prediction discrepancies (pd) [2]
  - normalised prediction distribution errors (npde) [3]
- Assumptions
  - model M^B has been built using a building dataset B
  - null hypothesis: this model can be used to describe the data collected in a validation dataset V (=B in internal evaluation)
- General class of Posterior Predictive Check (PPC), born in the Bayesian world
  - model M^B used to simulate data according to the design of V
  - compare a statistic computed on the real data in V to the distribution of the statistic obtained through the simulations
  - here *plug-in* approach (ignoring uncertainty)

Model and data

Statistical models

Model for observation y_{ij}

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

where:

- subject i ($i = 1,\ldots$N), with n_i observations $\mathbf{y}_i = \{y_{i1},\ldots,y_{in_i}\}$ at times t_{ij}, and covariates $\mathbf{z}_i$
- individual parameters θ_i
  - often modelled parametrically as a function h of fixed effects μ and random effects η_i:

$$\theta_i = h(\mu(\mathbf{z}_i), \eta_i) \text{ where } \eta \sim \mathcal{N}(0, \Omega)$$

  - in PK/PD, h is frequently a log-normal transformation, such that for the p$^{\text{th}}$ component:

$$\theta_{i(p)} = \mu_{(p)}(\mathbf{z}_i)\, e^{\eta_{i(p)}}$$

- f: structural model, common to all subjects
- g: residual error model, potentially depending on additional parameters, for instance

$$g(\theta_i, x_{ij}, \mathbf{z}_i) = a + b\ f^c(\theta_i, x_{ij}, \mathbf{z}_i) \qquad \text{(combined error model)}$$

Illustrative example

Dataset from 12 subjects given a single oral dose of theophylline used as a template to simulate illustrative datasets:

- 11 blood samples over a period of 25 hours (data at t=0 was omitted from the dataset for all patients): nominal times 15 and 30 min, 1, 2, 4, 5, 7, 9, 12, 24 h
- one-compartment model with first-order absorption
- variability models: IIV modelled using an exponential model, and combined error model for the residual variability

Table 1: *parameters estimated in original dataset*

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

Simulated datasets (N=100)

- V_{true}: simulated under M_B (H_0)
- V_{bioavail}: bioavailability divided by 2 (⇔ V/F multiplied by 2)
- V_{IIV}: IIV increased by 50% for V
- V_{2cpt}: simulated with a two-compartment model
  - k_a=1.55 hr^{-1}, V=20 L, k=0.02 hr^{-1}, k_{12}=0.2 hr^{-1}, k_{12}=0.01 hr^{-1}
  - 30% IIV on k_{12} and k_{12}
  - parameters re-estimated with a one-compartment model

Methods

Simulation-based metrics

Visual Predictive Check:

- K datasets $V^{sim(k)}$ simulated under model M^B using the design of the validation dataset V ($\mathbf{y}_i^{sim(k)}$: vector of simulated observations for the i^{th} subject in the k^{th} simulation)
- plot prediction interval corresponding to a given value (eg 90, 95%)

Prediction discrepancies and prediction distribution errors:

- F_{ij}: cumulative distribution function (cdf) of the predictive distribution of Y_{ij} under model M^B
  - F_{ij} obtained using Monte-Carlo simulations (same as VPC)
- prediction discrepancy for observation y_{ij}

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

  - where $\delta_{ijk} = 1$ if $y_{ij}^{sim(k)} < y_{ij}$ and 0 otherwise
  - pd expected to follow $\mathcal{U}(0,1)$ under the model
  - within-subject correlations introduced when multiple observations are available for each subject [2]
- prediction distribution errors
  - decorrelation using empirical mean $E_{\text{emp}\,i}$ and empirical variance-covariance matrix var($\mathbf{y}_i$) over the K simulations for simulated and observed data:

$$\mathbf{y}_i^{sim(k)*} = \mathbf{V}_{\text{emp}\,i}^{-1/2}(\mathbf{y}_i^{sim(k)} - E_{\text{emp}\,i})$$
$$\mathbf{y}_i^{*} = \mathbf{V}_{\text{emp}\,i}^{-1/2}(\mathbf{y}_i - E_{\text{emp}\,i})$$

  - pde obtained using decorrelated values and transformed to a normal distribution using the inverse of the normal cdf

$$\text{pde}_{ij} = F_{ij}^{*}(y_{ij}^{*}) \approx \frac{1}{K}\sum_{k=1}^{K}\delta_{ijk}^{*}$$
$$\text{npde}_{ij} = \Phi^{-1}(\text{pde}_{ij}) \qquad \sim \mathcal{N}(0,1) \text{ under } H_0$$

Graphs and tests

- Tests
  - VPC: no test (graphical approach), use Numerical Predictive Check
    - ∗ PI-NPC: compare percentages of outliers outside several prediction intervals to the theoretical value
  - pd and npde
    - ∗ Kolmogorov-Smirnov test: omnibus test
    - ∗ specific tests (Wilcoxon test for mean, Fisher test for variance, Shapiro-Wilks for normality), combined as a global p-value through a Bonferroni correction [3]
  - type I error inflation for non-corrected metrics induced by within-subject correlations [4]
- Graphs
  - VPC: visual diagnostic
  - the distribution of pd and npde can be assessed based on similar graphs as traditional residuals (eg WRES)
    - ∗ residuals versus time and predictions
    - ∗ histogram and QQ-plots
  - prediction bands around selected percentiles (obtained through repeated simulations under M^B) can be added to the different graphs

Results

Tests

- Simulations
  - performed under model M_B for the first three datasets
  - performed with 2-cpt model with parameters estimated
- Most tests detect the simulated model misspecifications, except:
  - KS test insensitive to IIV change
  - PI-NPC test on 80% interval insensitive to structural model misspecification

Dataset	Separate tests			Global tests		PI-NPC
	Mean	Variance	Normality	3 tests combined	KS test	80% PI
V_{true}	0.23	0.71	0.57	0.69	0.46	0.53
V_{bioavail}	$<10^{-9}$	0.002	$<10^{-10}$	$<10^{-10}$	$<10^{-15}$	$<10^{-15}$
V_{IIV}	0.78	0.01	0.69	0.04	0.51	4.10^{-6}
V_{2cpt}	0.001	0.79	0.64	0.002	0.005	0.11

Table 2: *Values of the tests on* npde *and of the binomial test on the coverage of the PI-NPC (90% PI), for the four datasets simulated in the present study.*

Graphs

Adding prediction bands and/or observed data may enhance the visual appeal of diagnostic graphs. Figure 1 shows an example with VPC:

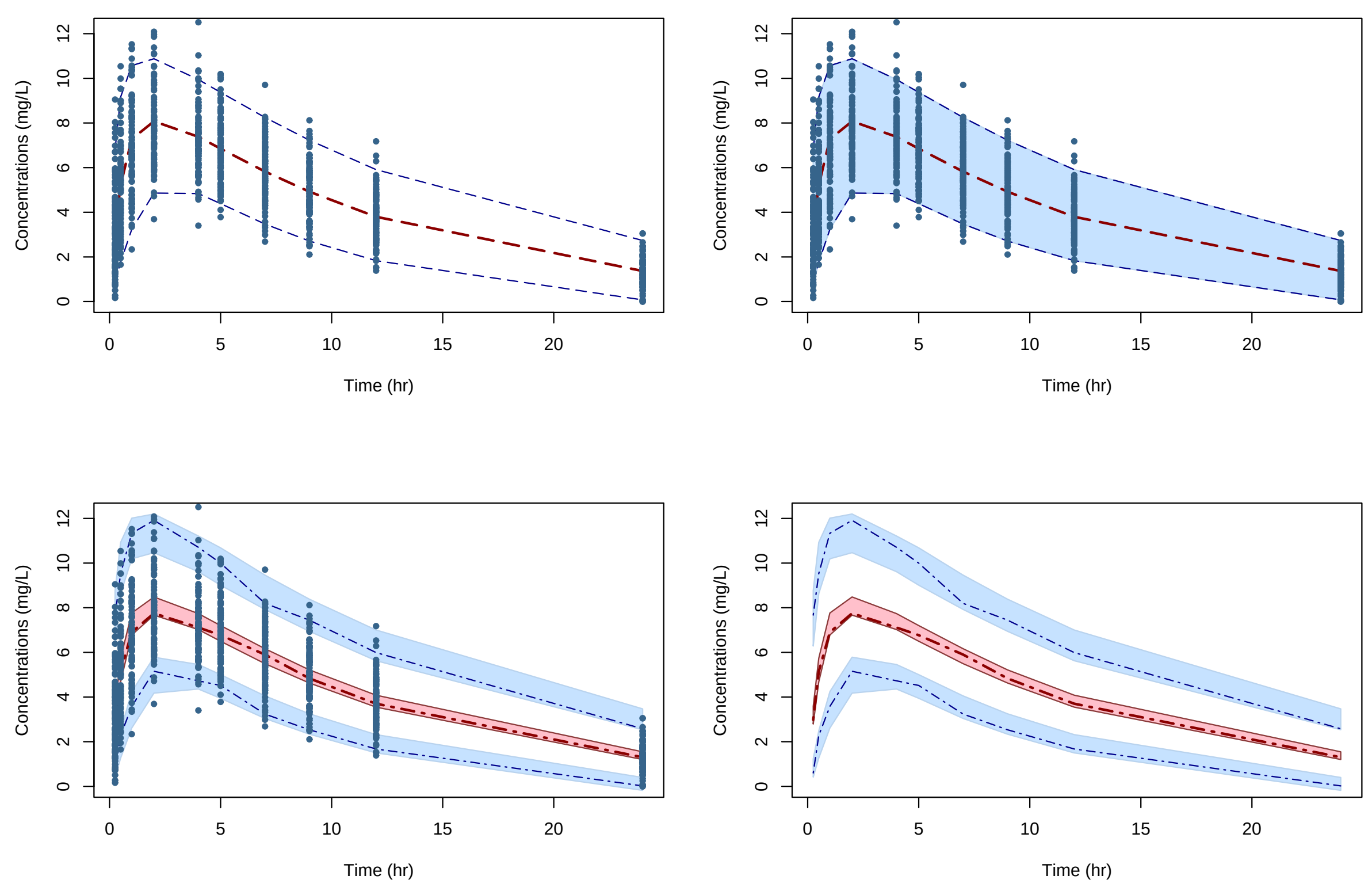

Figure 1: *VPC plots for V_{true}, with several representations. Top: 2.5 and 97.5$^{\text{th}}$ percentiles of the simulated data; thick dashed lines: 50$^{\text{th}}$ percentile; dots: observations. Bottom: 95% prediction intervals around 2.5, 50 and 97.5$^{\text{th}}$ percentiles (coloured areas); dotted/dashed lines: 2.5, 50 and 97.5$^{\text{th}}$ percentiles of observed data (thick line: median).*

Figures 2 and 3 show plots of VPC and pd versus time with prediction bands for the 4 simulated datasets.

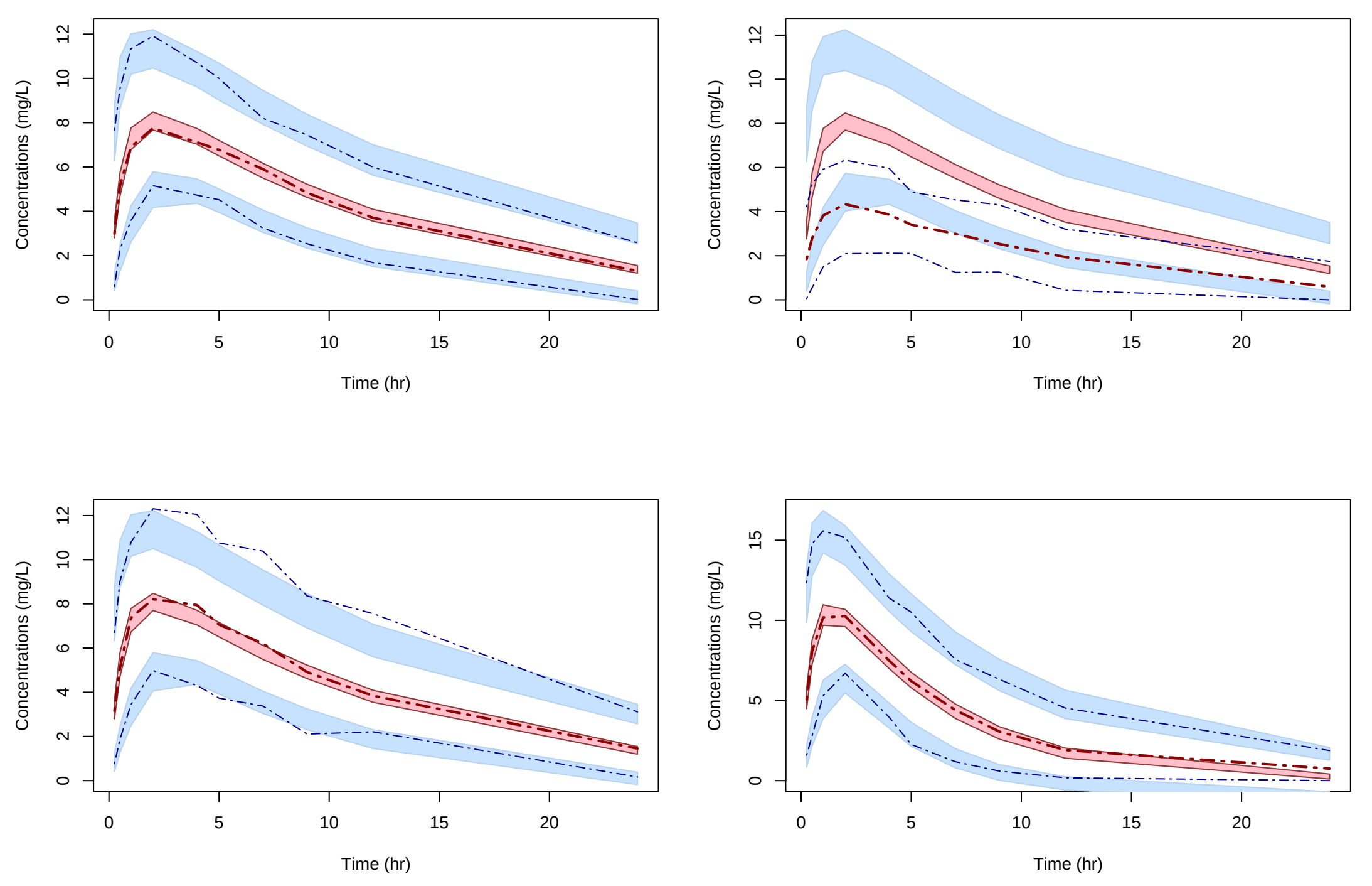

Figure 2: *95% VPC with prediction bands, for datasets V_{true} (upper left), V_{bioavail} (upper right), V_{IIV} (lower left), V_{2cpt} (lower right).*

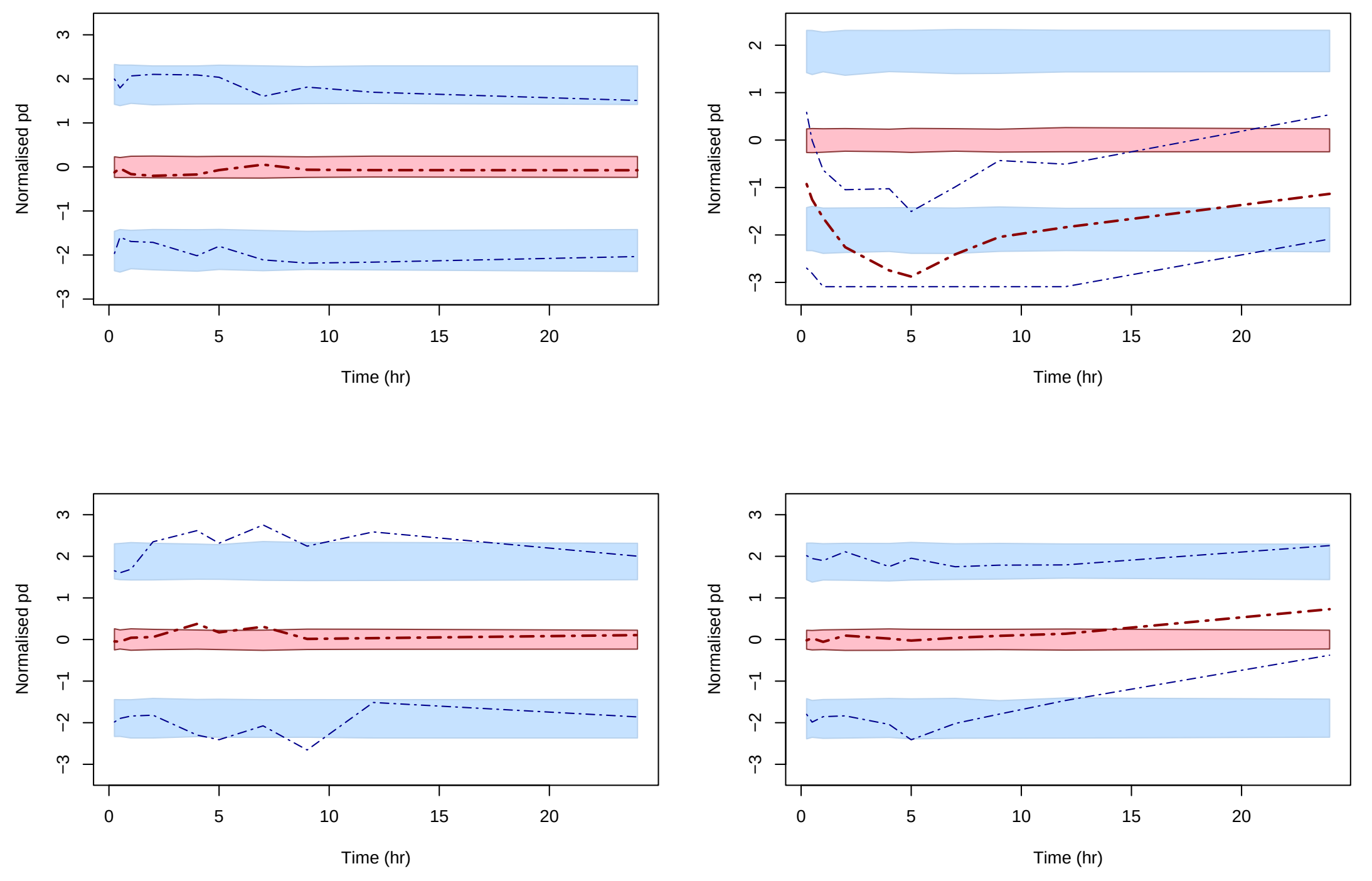

Figure 3: *Plot of pd versus time with prediction bands, for datasets V_{true} (upper left), V_{bioavail} (upper right), V_{IIV} (lower left), V_{2cpt} (lower right).*

Conclusion

- Array of complementary tools to be used by modellers
  - pd and VPC allow to visualise patterns with time
  - npde and PI-NPC provide a test
- Simulation-based metrics
  - require simulations under the model, which can be difficult to obtain, eg in the presence of drop-outs or censored data [5]
- Prediction bands obtained through repeated simulations
  - computer-intensive: final models only
  - enhance the detection model misspecifications by providing clear visual comparison of model expected behaviour versus observed data
- Tests
  - only npde provide adequate type I error thanks to decorrelation [4]
  - in real data, tests may be sensitive to large datasets or outliers
  - global tests: may be difficult to pinpoint exactly which aspects of the model to change
  - best used as a signal to guide further model improvement

References

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

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

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

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

[5] MO Karlsson and RM Savic. Diagnosing model diagnostics. *Clinical Pharmacology & Therapeutics*, 82:

★ Presenting author
email: emmanuelle.comets@inserm.fr