Development and performance of npde for the evaluation of time-to-event models

Marc Cerou, Marc Lavielle, Karl Brendel, Marylore Chenel, Emmanuelle Comets

Development and performance of npde for the evaluation of time-to-event models

M. Cerou *[1,2,3,4,5], M. Lavielle[6], K. Brendel [5], M. Chenel [5] and E. Comets[1,2,3,4]

[1]INSERM, CIC 1414, 35700 Rennes, France

[2]Université Rennes-1, 35700 Rennes, France

[3]INSERM, IAME, UMR 1137, F-75018 Paris, France

[4]Université Paris Diderot, IAME, UMR 1137, Sorbonne Paris Cité, F-75018 Paris, France

[5]Division of Clinical Pharmacokinetics and Pharmacometrics, Institut de Recherches Internationales Servier, Suresnes, France

[6]Inria & CMAP team Xpop, Ecole Polytechnique, University Paris-Saclay, Paris, France

*Corresponding author: marc.cerou@inserm.fr

Abstract

Purpose: Normalised prediction distribution errors (*npde*) are used to graphically and statistically evaluate mixed-effect models for continuous responses. In this study, our aim was to extend *npde* to time-to-event (TTE) models and evaluate their performance.

Methods: Let V denote a dataset with censored TTE observations. The null hypothesis (H_0) is that observations in V can be described by model $\mathcal{M}$. We extended *npde* to TTE models using imputations to take into account censoring. We then evaluated their performance in terms of type I error and power to detect model misspecifications for TTE data by means of a simulation study with different sample sizes.

Results: Type I error was found to be close to the expected 5% significance level for all sample sizes tested. The *npde* were able to detect misspecifications in the baseline hazard as well as in the link between the longitudinal variable and the survival function. The ability to detect model misspecifications increased as the difference in the shape of the survival function became more apparent. As expected, the power also increased as the sample size increased. Imputing the censored events tended to decrease the percentage of rejections.

Conclusions: We have shown that *npde* can be readily extended to TTE data and that they perform well with an adequate type I error.

Keywords— Model evaluation, npde, time-to-event, PSA

Abbreviations

BLQ: Below the limit of quantification

IIV: Inter-individual variability

KM: Kaplan Meier

KMVPC: Kaplan Meier visual predictive check

NLMEM: Nonlinear mixed-effect models

npde: normalised prediction distribution errors

pd: prediction discrepancies

PSA: Prostate specific antigen

TTE: Time to event

VPC: Visual predictive check

Introduction

An aim of many clinical trials is to evaluate a difference in response between several treatment groups. Survival analysis or, more generally, time-to-event (TTE) analysis - as the event can be something other than death, is a growing topic in this field because the time to an event is a meaningful and interpretable measure of many efficacy and safety endpoints and is usually more informative than a simple yes/no response [1]. Examples of such endpoints are time to death, time until organ rejection, time until infection and time until a satisfactory response is achieved. TTE analysis can be applied to a single group of patients or subjects, or be used to compare the experience of different groups of patients or subjects [2]. It is also used in observational trials to determine and test the existence of an epidemiological association [3, 4]. Clinical trials focusing on survival time or on responses that evolve over time generally involve regular visits, at which multiple follow-up measurements are collected. With this design, a change in the terminal outcome measurement can be associated with a change in the exposure condition. Subjective measures can also be helpful in understanding the impact of the treatment or intervention, and recording information on how a patient feels during and after treatment is becoming increasingly common in clinical trials. Nonlinear mixed-effect models (NLMEM) have become an integral part of drug development. They are widely used to capture both intra- and between-subject variabilities, which characterise the longitudinal measurements used in population pharmacokinetic and pharmacodynamic analysis, and can take into account missing values or unbalanced data [5]. There has also been a recent growing interest in joint models [6], which extend NLMEM by linking the longitudinal trajectory of a variable (such as a biomarker) to survival, and adequate estimation methods have been proposed [7]. These models form part of the discipline of pharmacometrics, which aims at identifying sources of variability and differences in drug efficacy and safety among population subgroups [8].

NLMEM development consists of building a structural and statistical model, estimating its parameters, and improving the model through selection and evaluation procedures. Model evaluation, or qualification, is described in the regulatory guidelines of various medicines agencies and is required as part of the model's development process [9, 10, 11]. How models should be evaluated depends on the intended purpose of the analysis, and how that evaluation should be reported has been described in a white paper for industry [12]. Model evaluation consists of checking how well a dataset can be described by the model [13, 14] and can be both internal and external. Internal evaluation takes place during the model-building process and consists of assessing fit or predictive ability by using the same data that is used to estimate parameters and select

models. External evaluation on the other hand uses independent data.

Many diagnostic tools have been proposed and described [15], including residual-based diagnostics such as weighted residuals (WRES) [16], or simulation based approaches using posterior predictive check [17]. For NLMEM with continuous data, such as biomarker concentrations, Mentré and Escolano developed a model evaluation tool called prediction discrepancies (*pd*), which takes into account the nonlinearity of the model function. They showed that this metric performs better than the WRES (also called standardised prediction errors or SPE) that were previously used [18] because it does not require linearisation of the model, unlike WRES or conditional WRES [19]. Brendel et al. developed a decorrelated version of the *pd*, called normalised prediction distribution error (*npde*) [20] in order to account for the correlation between multiple observations within an individual, and proposed a test to compare the distribution of calculated *npde* and their theoretical distribution. They also demonstrated good performance of *npde* through both statistical tests and graphical diagnostics [21, 22]. Both *pd* and *npde* can handle heterogeneous designs without the need to stratify on covariates or dosing regimens, offering an advantage over visual predictive checks (VPC), which are often used for their ability to show the evolution of the process being modelled [23, 24]. *npde* have recently been extended to pharmacokinetics data below the limit of quantification (BLQ) through imputation approaches [25]. This tool is available in an add-on package for R [26] and *npde* computation for continuous data is now integrated in the main software [16, 27] used in pharmacokinetic/pharmacodynamic data analysis. *npde* has applications in both internal and external evaluation [21].

The purpose of this study was to extend *pd* and *npde* to parametric TTE models by generalising the approach previously developed to handle BLQ data for different types of censoring. In that approach, prediction discrepancies for non-observed data were imputed within their expected distribution under the null hypothesis that the model is appropriate to generate the data. In the Models and methods section of this paper, we present the construction of *pd*/*npde* for TTE data. Then, by means of an extensive simulation study, we evaluate the performance of *npde* in terms of type I error and power to detect model misspecification in the context of external evaluation.

Models and methods

Statistical models

npde were first developed for continuous longitudinal responses. In this study, our objective was to extend this metric in the context of TTE models in order to address the TTE component of joint models. We therefore considered a model that would describe the impact of the continuous evolution of a biomarker on the risk of an event.

Time-to-event model

In clinical trials involving survival outcomes, patients are typically considered to be at risk of an event at time t. In this study, we modelled this risk, as opposed to non-parametric analyses where the focus of interest may be on difference in survival profile depending on covariates of interest.

Let T denote the variable representing the time of occurrence of the event, relative to a reference $t = 0$, usually the time when the patient enters the trial. Patients are generally observed for a limited period of time in a trial, during which they may or may not present the event of interest. The duration of the trial acts as a censoring mechanism and if the event does not occur within this period, the event is said to be *right-censored*. Other types of censoring can occur in clinical trials. Sometimes, the exact time of the event is not observed directly but the event is known to have occurred within a known time interval, in which case the event is called *interval-censored*. More rarely, in some specific designs, the true value of the event time is missing, but is known to be smaller than the observation time. In this case the event is *left-censored*.

Let T_i be the observation of the outcome in subject i. In standard survival analysis, T_i is associated with a variable C_i representing censoring. When C_i equals 0, the event is observed and T_i represents the time to the event, whereas when C_i equals 1, the event is censored and T_i is set to the censoring time. For interval-censored data, T_i is associated with a known interval $[T_{L_i}, T_{R_i}]$. Right or left censoring can be considered as a particular case of interval censoring, with respectively $T_{R_i} = +\infty$ or $T_{L_i} = 0$. If $T_{L_i} = T_{R_i}$ $(=T_i)$, then T_i is not censored. Regardless of how event times are recorded, we can define the survival function S, where, for any time $t \geq 0$, $S(t)$ is the probability of being event free until time t:

$$S(t) = Pr(T \geq t) \tag{1}$$

and where $S(0) = 1$.

An alternative is to use the hazard function, denoted by $h(t)$, which describes the instantaneous risk of having an event at time t for an individual having survived up until then. This can be expressed as:

$$h(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{dt} \tag{2}$$

By integrating the hazard from 0 to t, we obtain the following relationship between $S(t)$ and $h(t)$:

$$S(t) = \exp\left(-\int_0^t h(x)dx\right) \tag{3}$$

Time-to-event model involving a biomarker

In this study, we considered the special case in which a continuous variable, such as a biomarker, varies over time, influencing the risk of an event. Let $f(t, \theta_i)$ denote the prediction of the biomarker at time t in subject i. We assumed that f is a known nonlinear function supposed to be identical for all individuals, depending on a vector of individual parameters θ_i. In NLMEM, the individual parameters θ_i are decomposed into fixed effects μ, representing typical effects of the population, and random effects η_i, specific to each individual. Following [14], we assumed a joint multinormal distribution for the vector of random effects η_i and that there exists a transformation g such that $g(\theta_i)$ can be expressed as a linear function of μ, η_i, and possibly covariates. Usual transformations in pharmacokinetics or pharmacodynamics include the identity function, which yields a normal distribution for the parameters, the logarithmic function, which yields a log-normal distribution, and the logit function, which yields a logit-normal distribution. The variance-covariance matrix Ω of the random effects quantifies the magnitude of interindividual variability (IIV), as each diagonal element ω_k^2 represents the variance of the kth component of the random effects vector. For the purposes of this work, we assumed that the structural model f and the parameters of the statistical model are known.

A joint model can be defined by considering the relationship between the hazard function h and the evolution of the biomarker f. A standard practice is to decompose h as follows:

$$h_i(t \,|\, f_i(t)) = h_0(t) \exp(\beta f_i(t, \theta_i)) \tag{4}$$

in which β represents the strength of the association between the longitudinal outcome and the instantaneous risk h. If $\beta = 0$, the hazard is not affected by the evolution of the marker, so there is no link between biomarker and survival. $h_0(t)$ denotes the baseline risk function. In our joint modelling framework, h_0 was a parametric function (as opposed to non-parametric or semi-parametric models) defined by a vector of parameters $\lambda = (\lambda_\ell, 1 \leq \ell \leq n_\lambda)$. Using a parametric model allowed us to define the joint likelihood function.

Note that because the event was unique for each individual, parameters λ and β were assumed to have no IIV for reasons of identifiability.

Finally, we call Ψ the vector of population parameters, where $\Psi = \{\mu, \Omega, \lambda, \beta\}$ for our purposes. For full joint models, we could also include the parameters of the measurement error model on the longitudinal outcome, as well as covariate effects.

npde for TTE models

Model evaluation for TTE response

In a previous paper, we proposed an approach for model evaluation of continuous responses suited to NLMEM, by defining residuals that take into account the nonlinearity in the structural model [20]. We tend to consider *npde* as residuals because although they are not defined as the difference between observations and predictions, their distribution can be interpreted in a similar way to the more traditional residuals, which suffer from poor statistical properties in NLMEM [18, 20]. Also, we can evaluate *npde* using the same graphs as with standard residuals. In this study, our objective was to extend these residuals to TTE data. Considering the time to the event rather than the yes/no occurrence of the event allows the same approach to be used as for continuous response. Denoting V as a dataset with TTE observations and $\mathcal{M}$ as the model to be tested, our null hypothesis (H_0) is that the observations in V can be described by model $\mathcal{M}$.

Residuals called *pd* were developed by Mentré and Escolano [18] and decorrelated in [20] to account for repeated measurements to obtain *npde*. *pd* are defined as the quantile of an observation within its predictive distribution. Let $p_i(T_i|\Psi)$ denote the predictive distribution of observation T_i in the individual i under the model $\mathcal{M}$ being tested. In NLMEM, we only know how to write the probability of an observation conditionally to the individual parameters θ_i, so the predictive distribution is obtained by integrating over the distribution of the random effects:

$$p_i(T_i|\Psi) = \int p(T_i|\theta_i, \Psi)p(\theta_i|\Psi)d\theta_i \tag{5}$$

Note that when the structural model is nonlinear, there is no analytical solution for the integral in (5). Denoting F_i as the cumulative predictive distribution function of T_i under the tested model, the prediction discrepancy pd_i for T_i is defined as the value of F_i at the observation T_i:

$$pd_i = F_i(T_i) = \int_0^{T_i} p_i(t|\Psi)dt = \int_0^{T_i} \int p(t|\theta_i, \Psi)p(\theta_i|\Psi)d\theta_i dt \tag{6}$$

Mentré and Escolano [18] showed that construction of the *pd* implies that they follow a uniform distri-

bution under H_0. A pd_i corresponds to the quantile of the observation in its predictive distribution. If the model is well characterised, with the variability adequately taken into account, these quantiles are expected to distribute uniformly over the interval [0,1].

Taking censoring into account

To address the censoring inherent to this type of response, we used the same imputation approach as Nguyen et al. [25].

In the general case of interval-censoring, where the observation T_i lies within an interval $[T_{L_i}, T_{R_i}]$, we computed the probability of being below T_{L_i} (respectively T_{R_i}) as:

$$P(T_i \leq T_{L_i}) = F(T_{L_i}) = \int_0^{T_{L_i}} p_i(t|\Psi)dt \tag{7}$$

pd_i lies within the interval $[F(T_{L_i}), F(T_{R_i})]$, and we proposed to impute it from a uniform distribution $U\left(F(T_{L_i}), F(T_{R_i})\right)$ as, under the model, the distribution of the pd is known to be uniform. This definition generalises to the right-censored case in which $F(T_{R_i}) = 1$ and to the left-censored case in which $F(T_{L_i}) = 0$. When the exact time of the event is observed, $F(T_{L_i}) = F(T_{R_i})$, meaning that no imputation is required as the usual definition holds.

Figure 1 illustrates how prediction discrepancies are derived from an observed event on the X-axis through the cumulative predictive distribution (solid line). The dashed lines show the interval over which pd are imputed in the case of interval-censored TTE.

As previously, by construction, the pd_i's follow a uniform distribution $\mathcal{U}(0,1)$ under H_0. pd_i's are generally transformed into a normal distribution using the inverse function of the cumulative distribution function ϕ of $\mathcal{N}(0,1)$:

$$npde_i = \phi^{-1}(pd_i) \tag{8}$$

In this study, we considered survival data so that only one TTE observation is available for each individual, and the resulting $npde$ are expected to follow a $\mathcal{N}(0,1)$ distribution under the null hypothesis that $\mathcal{M}$ describes adequately the data in V [21].

Test based on npde

To compare the $npde_i$ with their theoretical distribution, we used the global test proposed in [20], which first involves performing three tests: (1) Wilcoxon signed rank test to check whether the median significantly

differs from 0; (2) Fisher test to check whether the variance significantly differs from 1; and (3) a Shapiro-Wilk test to check whether the distribution significantly differs from a normal distribution. As part of the global test, these three tests were then combined with a Bonferroni correction for multiple comparisons: for a given significance level α, the null hypothesis was rejected if one of the three p-values was smaller than $\alpha/3$.

Practical implementation

In NLMEM, the predictive distribution $p_i(T_i|\Psi)$ given by Equation 5 has no analytical expression. Mentré and Escolano proposed to approximate it using Monte Carlo simulations [18]. As such, K datasets were simulated to approximate the predictive distribution of each observation. K must be chosen large enough to provide a good approximation [26].

Using the approximated predictive distribution, the probability of being below T_{L_i} (resp. T_{R_i}) was estimated for each individual i from the model as the fraction of simulated values $T_i^{sim(k)}$ (k = 1, $\cdots$, K) smaller or equal to T_{L_i} (resp. T_{R_i}):

$$\widehat{Pr}\left(T_i \leq T_{L_i}\right) = \hat{F}_i(T_{L_i}) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}_{T_i^{sim(k)} \leq T_{L_i}} \tag{9}$$

in which $\mathbb{1}_c$ equaled 1 if condition c was true and 0 otherwise. When $T_i < T_i^{sim(k)}$ for all replicates k (k=1,$\cdots$,K), $pd_i = 0$ and similarly when $T_i > T_i^{sim(k)}$ for all replicates k, $pd_i = 1$. However, setting pd_i to 0 or 1 created numerical problems when converting to *npde* with eq (8). Therefore, in the case in which $T_i < T_i^{sim(k)}$ for all k, pd_i was set to a random sample from a uniform distribution between 0 and $\frac{1}{K}$. Conversely when $T_i > T_i^{sim(k)}$ for all k, pd_i was sampled from a uniform distribution between $1 - \frac{1}{K}$ and 1.

In practice, and especially if K was small, two observed values may have fallen whithin the same quantile of their respective predictive distributions, leading to ties in the *pd* and *npde* distribution. To avoid this, we chose to resample *pd* within the uniform distribution defined by the two consecutive quantiles surrounding the *pd* $\mathcal{U}(pd, pd + \frac{1}{K})$.

Graphs

pd and *npde* provide valuable tools for graphical diagnostics, enabling the detection of various model mis-specifications [22]. For TTE, we were able to compare the empirical distributions with their theoretical counterparts through quantile-quantile plots or histograms to assess when the model overpredicts or under-predicts event occurrence. For instance, observed events occurring later than predicted are materialised as a shift to the right of the median value in the histogram of *npde*. For continuous data, we were able to

compute individual predictions allowing us to look at trends versus time or predicted values. With single TTE data however, the outcome was confounded with time itself, meaning that such scatterplots were not as informative.

Evaluating npde performance for TTE data

In the previous section, we showed how to extend *npde* to TTE data. We evaluated their performance using a simulation study based on a joint model linking the evolution of a biomarker with survival in prostate cancer patients [28]. In this section, we give a brief description of Motivating example and underlying models, followed by the simulation settings used to evaluate the performance of *npde* to detect various model misspecifications in the risk function. In this study, we considered external evaluation of the TTE component alone, with two key assumptions: (i) we did not have observations of the biomarker, only those of the time to survival; and (ii) we assumed that the model for the biomarker was correct.

Motivating example

Prostate cancer is the most common form of cancer in men and the second leading cause of death from cancer in developed countries [29]. For metastatic castration-resistant prostate cancer, evaluation of treatment efficacy relies primarily on overall survival. The patients are also monitored throughout the trial by measuring the prostate-specific antigen (PSA): since cancer cells produce PSA, the level of PSA is an indicator of tumour size, which is related to overall survival. Therefore, PSA can be used as a surrogate biomarker in cases of declared prostate cancer [30, 31]. In a healthy adult male population, the PSA value ranges from 2.5 (at 50 years of age) to 6.5 ng/mL (at 80 years of age), depending on age and number of prostate cells. In a study by Desmée et al. [28], $N = 500$ patients were considered with a maximum follow-up of 735 days. Measurement was every 21 days leading to a maximum of 36 PSA observations per patient, with only death considered as a drop-out mechanism.

For the TTE model in our study, we used that developed by Desmée et al, who proposed a joint model for PSA kinetics and survival data from the VENICE trial [32]. The model for the hazard built by Desmée et al. relates the current PSA prediction to the hazard function:

$$h(t|PSA(t,\theta)) = h_0(t)\exp(\beta \times PSA(t,\theta)) \tag{10}$$

in which β corresponds to the strength of the link between PSA and the risk of having an event. h_0 is

described by the following parametric Weibull model:

$$h_0(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \tag{11}$$

in which k represents the shape and λ the scale in this model.

For the longitudinal model, PSA (ng.mL^{-1}) is assumed to be secreted by all prostate cells C (mL^{-1}), regardless of whether these cells are cancerous. In the absence of treatment, prostate cells proliferate at rate r (day^{-1}) and are eliminated at rate k_{out} (day^{-1}). They secrete PSA at a production rate p (ng.day^{-1}), and the antigen is cleared at rate δ (day^{-1}). Treatment consists of chemotherapy for metastatic castration-resistant prostate cancer, and the treatment is assumed to block cell proliferation with time-varying effectiveness, due to the onset of cancer cell resistance. The proliferation rate under treatment is given by $r' = r(1-e(t))$ with $e(t)$ in [0,1] representing time-dependent treatment effect. Figure 2 depicts the evolution of PSA secreted by prostate cells and it can be described by ordinary differential equations:

$$\begin{cases} \frac{dC(t)}{dt} &= r(1-e(t))C(t) - k_{out}C(t) \\ \frac{dPSA(t)}{dt} &= pC(t) - \delta PSA(t) \end{cases} \tag{12}$$

Treatment is supposed to have a constant efficacy (ϵ) until a time T_{esc}, when it becomes ineffective, allowing the tumour to proliferate:

$$e(t) = \begin{cases} \epsilon & \text{if} \quad t \leq T_{esc} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

A quasi steady-state assumption at treatment initiation is made, allowing to estimate prostate cells and cancer cells at baseline $C_0 = \frac{\delta \times PSA_0}{p}$ in which PSA_0 is the value of PSA at baseline. With a piecewise constant treatment effect, the model has an analytical solution:

$$PSA(t,\theta) = \begin{cases} \frac{\delta PSA_0}{r(1-\epsilon)-k_{out}+\delta} e^{(r(1-\epsilon)-k_{out})t} + \left[PSA_0 - \frac{\delta PSA_0}{r(1-\epsilon)-k_{out}+\delta}\right] e^{-\delta t} & \text{if} \quad t \leq T_{esc} \\ \frac{\delta PSA_0}{r-k_{out}+\delta} e^{(r-k_{out})t - r\epsilon T_{esc}} + \left[PSA(T_{esc}) - \frac{\delta PSA_0 e^{(r(1-\epsilon)-k_{out})T_{esc}}}{r-k_{out}+\delta}\right] e^{-\delta(t-T_{esc})} & \text{otherwise} \end{cases} \tag{14}$$

Table I shows the population parameters estimated by Desmée et al. for this model on the VENICE data [28], rounded for the purposes of the present simulation study. In [28], k_{out} and δ were set respectively at 0.046 day^{-1} and 0.23 day^{-1}, owing to issues with the identification of model parameters. IIV was estimated for four parameters, $[r, PSA_0, \epsilon, T_{esc}]$, with a logit distribution for the treatment effect ϵ and a log-normal distribution for the three others. Ω was assumed to be diagonal, with the last column in Table I giving the standard deviation of the random effects. There was no variability on the parameters of the TTE model (Equation 10).

Table I: Values of the population parameters for PSA used for the simulation in all scenarios, and parameters of the base Weibull model for the TTE component.

Parameter	Fixed effects	Transformation	Inter-individual standard deviation
r	0.05	log-normal	0.1
PSA_0	80	log-normal	0.6
ϵ	0.3	logit-normal	1.5
T_{esc}	140	log-normal	0.6
k	1.5	-	0
λ	580	-	0
β	0.001	-	0

Simulation study

The objective of the simulation study was to evaluate the performance of *npde* in terms of type I error and power for different types of model misspecifications. The general evaluation framework is described in Figure 3. For each scenario, 1) the dataset V was generated under model $\mathcal{M}_V$, and 2) using Monte Carlo simulations, K replicates based on the same design as V were simulated under the tested model $\mathcal{M}$ to approximate the predictive distribution. This process was repeated a number of times (B=200), and then 3) the statistical performance of *npde* was assessed as the empirical probability of rejecting the hypothesis that $\mathcal{M}$ describes V.

In the first simulation, we evaluated the performance of *npde* to detect misspecification in the link function, specifically through the exponential term of Equation 10 representing the impact of the longitudinal variable on survival. TTE data for V were simulated according to one of two models: (1) the base model $\mathcal{M}_{V,medium}$ with $\beta = 1e-3$ (which represents an arbitrary moderate effect of PSA on survival) and parameters from Table I, or (2) a model $\mathcal{M}_{V,noLink}$ with $\beta = 0$ (no effect of PSA on survival).

For each of these two settings, we then tested four models with the same baseline hazard and β equal to 0, 1e-4, 1e-3 or 5e-3, corresponding respectively to models without a link ($\mathcal{M}_{noLink}$), or with a low ($\mathcal{M}_{low}$), medium ($\mathcal{M}_{medium}$) or high link ($\mathcal{M}_{high}$). These tested models were used to build the predictive distribution and to compute *pd* and *npde*.

Each scenario (defined as a combination of simulated data and tested model) was run with four different sample sizes for the study, $N = \{50, 100, 250, 500\}$.

The aim of the second simulation was to check the performance of *npde* in detecting misspecification in the baseline risk h_0 of Equation 10. In this scenario, V was again simulated according to the model described in Table I.

We tested five models for $\mathcal{M}$ with β and λ unchanged, and k equal to 1, 1.3, 1.5, 1.7 and 2. Note that the case in which k equals 1 is the exponential distribution with constant hazard function. Each scenario was run with three values of sample size for the study, $N = \{50, 250, 500\}$.

Simulation settings

In all scenarios, PSA was simulated as described in the Motivating example with the parameters in Table I both under H_0 and H_1, so that no model misspecification was assumed for the longitudinal component. We did not simulate any PSA data as we only needed the predictions of the model to approximate the predictive distribution for the TTE component of the model. This corresponded to an external evaluation in late-phase trials in which only observations of time to death are recorded. We evaluated the performance of *npde* both using the full data (no censoring) and assuming a follow-up duration of $t = 735$ days. Event times were assumed to be observed during the trial and only right-censoring was considered.

For each scenario and each candidate model $\mathcal{M}$, the predictive distribution was computed once, using $K = 1000$ datasets simulated under $\mathcal{M}$. This predictive distribution was then used to compute the *pd*/*npde* for the 200 datasets V. The proportion of simulated datasets in which $\mathcal{M}$ was rejected was defined as the type I error for scenarios in which $\mathcal{M}_V$ and $\mathcal{M}$ were the same (H_0), whereas it corresponded to the power for scenarios in which $\mathcal{M}_V$ and $\mathcal{M}$ were different (H_1):

$$\text{Type I error} = \frac{\#\text{ of rejected datasets simulated under } H_0}{\#\text{ of datasets simulated under } H_0} \tag{15}$$

and

$$\text{Power} = \frac{\#\text{ of rejected datasets simulated under } H_1}{\#\text{ of datasets simulated under } H_1} \tag{16}$$

Assuming a theoretical p-value of 5% for the global adjusted test, we computed the expected prediction interval for the 200 tests using the exact Binomial test as [0.024-0.09].

Implementation

We used the statistical software R [33], version 3.2.3, to implement the computation of the *pd* and *npde* for TTE data and to perform the statistical tests. We also simulated data in R for the simulation study,

by predicting PSA kinetics with the analytical function (14) and using the *simulx* function from the *mlxR* package to simulate TTE data. An R script with the code for the simulation of the events is provided in the Supplementary material.

Results

Data

Figure 4 shows the predicted evolution of the logarithm of the PSA for each individual over time for one dataset with the base model described in the 'Motivating example' (see Table I), with $N = 500$ subjects. To represent the observations of the time to survival, we show the associated Kaplan Meier (KM) curve as a solid line. In this particular scenario, there is a "medium" link between PSA kinetic and survival ($\mathcal{M}_{V,medium}$) which is why very high PSA values (over 4000) are not observed, as they would be associated with patients death.

Graphical diagnostics using npde

We computed the *npde* for the dataset simulated in the previous section, under two tested models $\mathcal{M}$, $\mathcal{M}_{medium}$ (H_0), the same model used to generate V, and model $\mathcal{M}_{noLink}$, in which $\beta = 0$ (H_1). Figure 5 shows the diagnostic plots obtained with $\mathcal{M}_{medium}$ on the left and $\mathcal{M}_{noLink}$ on the right: quantile-quantile plots are shown in the top row and histograms in the middle row. Another way to evaluate the model is to look at the Kaplan Meier visual predictive check (KMVPC) plots which are represented in the bottom row. In these graphs, the survival in the data is estimated along with its 90% prediction interval based on the 1000 Monte Carlo simulation. Table II shows the corresponding p-values for the tests comparing each distribution to a theoretical normal distribution.

Table II: p-values of the tests performed on *npde* in cases under H_0 and H_1.

Case	Test of median	Test of variance	Normality test	Adjusted p-value
$H_0 : \mathcal{M} = \mathcal{M}_V$	0.267	0.483	0.585	0.799
$H_1 : \mathcal{M} \neq \mathcal{M}_V$	$< 10^{-5}$	0.055	0.363	$< 10^{-4}$

Under H_0, $\mathcal{M}_V$ and $\mathcal{M}$ correspond to model $\mathcal{M}_{medium}$ ($\beta = 0.001$). Under H_1, $\mathcal{M}_V$ corresponds to $\mathcal{M}_{medium}$ ($\beta = 0.001$) whereas $\mathcal{M}$ is $\mathcal{M}_{noLink}$ ($\beta = 0$)

In Figure 5 under H_0, the graphs show no major departures from the distribution of *npde* when compared

with the theoretical standard normal distribution, and the adjusted p-value of the test equals 0.799. Under H_1, the strength of PSA impact on survival in $\mathcal{M}$ is lower than in the model $\mathcal{M}_{V,medium}$ used to simulate V, leading to a longer average survival. This can be seen in the VPC graph as an overprediction of the survival. This also shows in the distribution of the *npde*: the histogram of the *npde* shows a negative shift of the median from the theoretical standard normal distribution leading to a rejection of the model. This is confirmed by the tests in Table *II*, in which the p-value of the median test is significant after Bonferroni correction, whereas the two other tests are not significant.

Misspecification of PSA impact on survival

In the first simulation, we focused on the impact of a misspecification in the link function by varying the parameter β. Figure 6 illustrates the performance of *npde* in terms of type I error and power. The curves show the proportion of simulated datasets V in which the tested model M was rejected, plotted against the value of β in M used to generate the distribution of *npde*. We tested two scenarios in which we investigated different strengths of the link and different sample sizes. The tested models are ordered on the X-axis with an increasing value of β and the proportion of rejection, which corresponds to the type I error or power, is on the Y-axis. Because this was a simulation study, true event times were known, meaning that we compared the performance both with (dashed line) and without (solid line) censoring.

In both scenarios, under H_0, the type I error falls within the prediction interval of the expected value (5%) for all sample sizes and regardless of whether the event times are censored. Censoring the data leads to a slight decrease in the type I error (numbers can be found in the additional Table SI given in the Supplementary material), but it remains within the variability materialised by the prediction interval shown in grey in Figure 6.

Considering the simulations under H_1, the power increases with the sample size, as expected, as well as with the difference between $\mathcal{M}_V$ and $\mathcal{M}$. That is to say, the difference between the "true" β and the one in the model $\mathcal{M}$ used to compute the predictive distribution. Larger β imply earlier occurrence of events, so that differences in β between M_V and M are reflected in a shift of the predictive distribution compared with observations in V. Models that overpredict the time to event will, by construction, induce a negative shift of the median of the *npde*. Accordingly under H_1, rejection of the global test is due in most cases to a rejection of the Wilcoxon test (data not shown), which tests whether the median of *npde* equals 0. However in the scenario in which $\mathcal{M}_V$ is $\mathcal{M}_{V,noLink}$ with $\beta = 0$ (left-hand figure), the rejection of the model is also

due to a rejection of the Fisher test (not shown) where the null hypothesis is that the variance equals 1.

Figure 6 shows that the proportion of rejected datasets decreases when the data are censored, leading to a decreased power under H_1. This is expected because in the presence of censored data, *pd* are imputed under the model being tested, so that part of the distribution of the *npde* is in fact imputed under H_0. We also observe that the loss of power is larger in the left-hand plots (V simulated under $M_{V,noLink}$) than in the right-hand plots (V simulated under $M_{V,medium}$), as shown by a larger difference between the dashed and solid lines. Part of the difference can be explained by the difference in the proportion of censored events in the two simulations. This proportion is around 25% for $M_{V,noLink}$, compared with around 16% for $M_{V,medium}$. As the *pd* for censored events are imputed in their expected distribution under the null hypothesis, we expect the p-value to be closer to the theoretical 5% with large amounts of censored events. To account for this difference, we adjusted the proportion of censoring by changing the censoring time to have the same proportion of censoring for the different settings. The results are shown in the Supplementary material (Figure S1) for three proportions of censoring (12, 25, and 50%); the figure shows that, as expected, a higher proportion of censoring led to a decrease in the power.

However, even after correcting for the difference in the proportion of censored events in the datasets, the loss of power remains higher in the scenario in which $M_V = M_{V,noLink}$. Looking at the distribution of the *npde*, we found a very skewed distribution in this case with a large difference between the censored and uncensored distribution at late times (data not shown), which can explain the larger loss of power under this model. This skewed distribution is due to late event times simulated under a model with no link, while the predictive distribution under the tested model in the alternatives H_1 is centred on earlier times. With the complete dataset, this translates to a group of *npde* clustered at the high end of the distribution, which is imputed to a uniform distribution under censoring. Additional simulations (see Table SI) in which V is simulated with a link β equal to 1e-4 (low link) or 0.005 (high link) show less loss of power when censoring data for the models in which biomarker evolution impacts survival through a non-null β, at least for moderate proportions of censoring (up to 25%).

Misspecification of the baseline hazard model

In the second simulation, we assessed the performance of *npde* to detect model misspecifications in the baseline hazard $h_0(t)$. Observed TTE data in V were simulated according to the base model $\mathcal{M}_{V,medium}$. The scenario under H_0 was the same as in the second scenario from the previous simulation, with a predictive

distribution obtained under the same model. We also computed the *npde* assuming four other Weibull models, by varying k between 1 (where the Weibull model is equivalent to an exponential distribution with constant hazard model) and 2, to evaluate the power under several alternatives H_1.

Figure 7 shows the proportion of datasets in which the model is rejected when the *npde* are computed under the five possible models.

This figure illustrates the ability of *npde* to detect misspecifications in h_0. Under H_0 (k=1.5), the type I error is close to 5% and remains whithin the prediction interval ([2.4%-9%]) both with and without censoring. As previously, the power increases as the difference between the true and tested value of k increases. Again, the power also decreases with a reduction of the sample size, dropping quite sharply below N=100. The proportion of simulated datasets in which the test is rejected is lower in the presence of censoring, which is consistent with an imputation under the model. Similar results were obtained with other scenarios under H_0 and are presented in Table SII of the Supplementary material.

Discussion

Joint models are increasingly used in clinical trials analysis, as they enable investigation of the relationship between a primary outcome, such as survival or the achievement of a response, and longitudinal variables, such as biomarkers, measured throughout the trial [34]. In previous studies, we developed a diagnostic approach, called *npde*, to evaluate NLMEM used to describe the evolution of continuous longitudinal variables [20, 21]. In this paper, we propose and evaluate an extension to *npde* that is adapted to the TTE data component of joint models.

npde are obtained by decorrelating and normalising the *pd*, defined by Mentré and Escolano as the quantile of an observation in its predictive distribution [18]. An advantage of *pd* and *npde*, compared with standardised residuals which involve a first-order approximation of the model function, is that their theoretical distribution is known. A global adjusted test was proposed [20], comparing the *npde* with a normal distribution with median 0 and variance 1. In a simulation study, the test was found to maintain a 5% type I error, in stark contrast with linearisation-based metrics which almost always rejected the model even under the null hypothesis [21]. *npde* also performed better than numerical predictive checks, which provide a statistical test for graphical VPC diagnostics, as they take into account the correlation intrinsically present in longitudinal data. In this study, we found that the *npde* also performed well for TTE models, maintaining an adequate

type I error in simulations evaluating misspecified baseline hazard or misspecified link function.

In our simulation study, only the survival model was evaluated, as we assumed that the evolution of PSA was similar in the evaluation dataset, and we did not simulate observations of the biomarker. Although the computation of *npde* involved predictions of the time course of the biomarker, it is worth noting that we did not need any individual observations of the longitudinal variable in our method, as we imputed the *pd* for censored events directly in the predictive distribution. This approach was first proposed to impute *pd* for BLQ data [25], and here we showed that it has a natural extension to right- or interval-censored data. For interval-censored data, we assumed that the event occurred within a predefined interval. An example of this would be to consider trials in which events were recorded at weekly or monthly visits. The exact time of the events was not known but they were supposed to have occurred since the last visit. We could surmise that imputing *pd* in this case could lead to loss of power, similar to what we observed for terminal censoring. However, provided the time intervals were not too large, we would still be able to detect shifts in median time to events even in the presence of censoring, so the loss of power would be limited. It would be interesting to investigate in a future study the impact of the duration of the interval, i.e. the impact on power if we were to consider weekly versus monthly visits, for example.

For continuous responses, an alternative method, used in Monolix [27], is to impute the censored observation to the model prediction. However, this method requires the definition of a predicted response, which is more difficult for survival data as we deal directly with the probability distribution. The two methods of imputation are available in the *npde* package [26] for continuous responses, but have not been formally compared. Both methods rely on the model, but imputing to the model prediction may be slightly more conservative when the residual variability is high and the proportion of censored data increases. This is because the resulting prediction discrepancies would ignore the residual error model, compared to imputing the *pd* directly within the cumulative distribution. It could be interesting to evaluate whether this approach could be extended for joint models in settings in which observations of the longitudinal biomarker are available, by defining the predicted time to event based on the individual survival function, which uses the information produced by the evolution of the biomarker [35].

The problem of defining predictions for TTE models also makes it more difficult to compute residuals, which are generally defined as the difference between observed and predicted data, possibly weighted by the expected variance of the prediction. As survival models are framed in terms of probability distributions, Cox-Snell residuals [36] have been proposed instead, which are computed using the logarithm of the estimated

survival function. Under H_0 and in the absence of censoring, Cox-Snell residuals follow an exponential distribution. They are generally used to check the overall fit of parametric TTE models [37]. Martingale residuals, on the other hand, are defined as the difference between the observed and expected number of events [38]. In contrast with the Cox-Snell residuals, they can be used both to evaluate model adequacy and to check the functional form for a covariate. However, their distribution is not known and tends to be asymmetric, so evaluation is mostly based on graphical diagnostics which may be subjective and difficult to interpret. Transformation to reduce their asymmetry leads to deviance residuals [39], which have been mostly used to help detect which individuals were poorly fit by the model, i.e. the outliers. None of these residuals have a known distribution under the null hypothesis except the Cox-Snell residuals, although the test needs to be adjusted in the presence of censoring, contrary to *npde*. We did not compare these metrics to *npde* in our study because the computations of all these residuals involve the estimation of individual parameters, which for the joint models would require data on the longitudinal biomarker, while in our simulation study we assumed that only TTE observations were available.

Simulation-based diagnostics can also be defined for TTE models, relying on the Bayesian idea of predictive distributions. KMVPC compares a nonparametric estimate of the survival function, obtained with a KM approach, with the survival prediction interval of the tested model. A second method has been proposed [40, 41], inspired by the KMVPC, which consists of comparing a nonparametric estimation of the hazard with a predictive interval under the tested model hazard. However, both approaches provide only a visual diagnostic, and there is no obvious way to derive a statistical test from the graphs, which makes it difficult to compare their performance with *npde*. Most of the time in our simulations, a misspecification seen in the VPC was associated with a significant p-value for the test on *npde*. However, in other cases under H_1, KMVPC was considered visually acceptable whereas *npde* detected a misspecification. Both *npde* and KMVPC can therefore be used as visual diagnostics, although *npde* also provides a statistical test.

Other tools for model evaluation in survival analysis have been described in the literature but the focus is more often on testing model components such as a covariate effect with, for example, Wald, Likelihood ratio and Score tests. The proportional hazard assumption can be assessed by plotting standardised Schoenfeld residuals versus time [42]. Also, to avoid making assumptions on the shape of the hazard, non-parametric or semi-parametric hazards are often used, and only sub-components of the model are tested.

The simulation study in this paper was based on a real example [28], and set in a context of external evaluation in which models were evaluated using data that was not used for model building. However, *npde*

are also used as an internal evaluation tool during model building, combined with Wald or Likelihhod ratio test used for model selection, in which case they can provide a visual diagnostic of model misspecifications making it possible to orient further model development [22]. In this case, the tested models M would be developed during the course of the analyses, and their parameters estimated instead of being fixed, as they were here, to arbitrary values. Using estimated parameters instead of theoretical values can be expected to have an impact on power as the estimation process may absorb some of the misspecification in the structural model. Simulation-based diagnostics are more computer-intensive, as the predictive distribution needs to be approximated by Monte-Carlo simulations in the absence of an analytical solution to compute the density in NLMEM. Following [15], we suggest computing *npde* for the final model and other key models in the analysis. Another issue in the computation of the predictive distribution is that the biomarker trajectory must be predicted to simulate event times. In this study, we used a simplified model with an analytical solution to predict PSA trajectories, but the use of ordinary differential equations can be much more time-consuming for complex models.

In our simulation settings, the power of the *npde* reached 100% in scenarios with a large difference in survival and a large number of subjects but, as expected, could be much lower when the tested model was close to the one generating the data. There is no gold standard against which this power can be compared, as there was no biomarker observations making it possible to fit the joint model in our simulations. This is both a strength and a weakness of the *npde*, in that we can compute *npde* in cases such as this where we cannot use the Likelihood ratio or the Wald tests but, in our setting, performance of the *npde* depended on a correct specification of the biomarker model. In our study, we assumed that only TTE observations were available, for instance, in the form of external evaluation data from a new study. Note also that there was no IIV here in the survival model, as a model with IIV would not be identifiable without repeated observations of the event. As a result, the predictive distribution of the biomarker was similar in V and M in all our scenarios, so that we effectively assessed only the fixed parameters associated with the survival function. Additional studies are needed to explore this issue further by considering repeated TTE models, as well as investigating misspecifications in the biomarker model and their impact on the power of *npde* for the TTE component. In a future study, we will investigate decorrelation methods for *npde* when considering multiple responses in the joint model, as well as the extension of *npde* to repeated TTE data, including IIV in the survival parameters.

We performed additional simulations with other distributions, such as that of Gompertz, to represent the baseline hazard (not shown). We found similar results in terms of type I error and a high power, as

these distributions induced marked changes in the time course of event occurrences. The type I error usually remained within the prediction interval expected for a theoretical p-value of 5%, except in one case shown in the Supplementary material (Table SII). The inflation in that case appeared to be due to sampling variability with an insufficient number of Monte Carlo samples used to build the predictive distribution. This may be an issue when there is a strong skewness in the distribution of TTE under some distributional assumptions. Indeed, very high values of event time could be simulated in V, for instance with model $\mathcal{M}_{noLink}$, without appearing in the predictive distribution (which was built only once within one scenario as it was a time-consuming step). These extreme values biased the estimate of the variance and because our test on *npde* includes a Fisher test to check if the variance of *npde* equals 1, the proportion of dataset which was rejected increased. On the other hand, observations lying outside of the predictive distribution can also mean that the model underestimates observations, so the test must be robust enough to reject the model when this is the case. The global test proposed in [20], combining a test of normality with a test of the median and a test of the variance through a Bonferroni correction, performed well with continuous responses in a simulation study [21], but the Bonferroni correction assumes that the three tests are independent which may not always be a valid assumption. Alternatives to this test could be envisaged, such as a global Kolmogorov-Smirnov test or tests on the distribution of the *pd*. In any case, the stability of the p-value for the test based on *npde* can be assessed by increasing the number of replications used to compute the predictive distribution, which should be performed especially for the final model to be evaluated. An estimate of the threshold for the test statistic can be obtained through a simulation study using the design of the trial, and can be used to correct the power if inflation is observed. In our study, no correction was used for the power reported in the results as the type I error remained within its prediction interval.

Conclusion

In conclusion, we propose an extension to *npde* making it possible to evaluate models for TTE data. This is an extension that performed well in a simulation study based on a joint model coupling biomarker evolution with time and survival, in which we investigated the type I error and power to detect model misspecification in the link between biomarker and survival as well as in the baseline hazard function.

Acknowledgements and Disclosures

Marc Cerou received funding from Institut de Recherches Internationales Servier. The authors thank Hervé Le Nagard and Francois Cohen for the use of the computer cluster services hosted on the "Centre de Biomodélisation UMR1137" and Solène Desmée for her help concerning the setup of the simulation study.

References

[1] Holford N. A time to event tutorial for pharmacometricians. CPT Pharmacometrics Syst Pharmacol. 2013;2:e43.

[2] Flynn R. Survival analysis. J Clin Nurs. 2012;21(19-20):2789–97.

[3] Versmissen J, Oosterveer DM, Yazdanpanah M, Defesche JC, Basart DCG, Liem AH, et al. Efficacy of statins in familial hypercholesterolaemia: a long term cohort study. BMJ. 2008;337:a2423.

[4] de Oliveira C, Watt R, Hamer M. Toothbrushing, inflammation, and risk of cardiovascular disease: results from Scottish Health Survey. BMJ. 2010;340:c2451.

[5] Mould D, Upton R. Basic concepts in population modeling, simulation, and model-based drug development. CPT Pharmacometrics Syst Pharmacol. 2012;1(9):1–14.

[6] Ibrahim J, Chu H, Chen L. Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. J Clin Oncol. 2010;28(16):2796–2801.

[7] Mbogning C, Bleakley K, Lavielle M. Joint modelling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the stochastic approximation expectation–maximization algorithm. J Stat Comput Simul. 2015;85(8):1512–28.

[8] Ette E, Williams P. Pharmacometrics: the science of quantitative pharmacology. Hoboken, New Jersey: John Wiley and Sons; 2013.

[9] Food and Drug Administration. Guidance for Industry Population Pharmacokinetics:Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER); 1999. Available from: http://www.fda.gov/downloads/Drugs/Guidances/UCM072137.pdf.

[10] Food and Drug Administration. Guidance for Industry Exposure-Response Relationships– Study Design, Data Analysis, and Regulatory Applications:Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER); 2003. Available from: https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm072109.pdf.

[11] Agency EM. Guideline on reporting the results of population pharmacokinetic analysis CHMP; 2007. Available from: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003067.pdf.

[12] Marshall S, Burghaus R, Cosson V, Cheung S, Chenel M, DellaPasqua O, et al. Good Practices in Model-Informed Drug Discovery and Development: Practice, Application, and Documentation. CPT Pharmacometrics Syst Pharmacol. 2016;5(3):93–122.

[13] Brendel K, Dartois C, Comets E, Lemmenuel-Diot A, Laveille C, Tranchand B, et al. Are population PK and/or PD models adequately evaluated? A 2002 to 2004 literature survey. Clin Pharmacokinet. 2007;46(3):221–234.

[14] Lavielle M. Mixed effects models for the population approach: models, tasks, methods and tools. London: Chapman and Hall/CRC Biostatistics Series; 2014.

[15] Nguyen T, Mouksassi MS, Holford N, Al-Huniti N, Freedman I, Hooker A, et al. Model Evaluation of Continuous Data Pharmacometric Models: Metrics and Graphics. CPT Pharmacometrics Syst Pharmacol. 2017;6(2):87–109.

[16] Beal S, Sheiner L, Boeckmann A, Bauer R. NONMEM Version 7.4. Ellicott City; 1989-2017.

[17] Yano Y, Beal SL, Sheiner LB. Evaluating pharmacokinetic/pharmacodynamic models using the posterior predictive check. J Pharmacokinet Pharmacodyn. 2001;28(2):171–92.

[18] Mentré F, Escolano S. Prediction discrepancies for the evaluation of nonlinear mixed-effects models. J Pharmacokinet Pharmacodyn. 2006;33(3):345–67.

[19] Hooker AC, Staatz CE, Karlsson MO. Conditional weighted residuals (CWRES): a model diagnostic for the FOCE method. Pharm Res. 2007;24(12):2187–2197.

[20] Brendel K, Comets E, Laffont C, Laveille C, Mentré F. Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. Pharm Res. 2006;23(9):2036–49.

[21] Brendel K, Comets E, Laffont C, Mentré F. Evaluation of different tests based on observations for external model evaluation of population analyses. J Pharmacokinet Pharmacodyn. 2010;37(1):49–65.

[22] Comets E, Brendel K, Mentré F. Model evaluation in nonlinear mixed effect models, with applications to pharmacokinetics. J Soc Fr Statistique. 2010;151(1):106–28.

[23] Holford N. The visual predictive check—superiority to standard diagnostic (Rorschach) plots. PAGE 14. 2005;Abstr 738. Available from: www.page-meeting.org/?abstract=738.

[24] Karlsson M, Holford N. A tutorial on Visual Predictive Checks. PAGE 17. 2008;Abstr 1434. Available from: www.page-meeting.org/?abstract=1434.

[25] Nguyen THT, Comets E, Mentré F. Extension of NPDE for evaluation of nonlinear mixed effect models in presence of data below the quantification limit with applications to HIV dynamic model. J Pharmacokinet Pharmacodyn. 2012;39(5):499–518.

[26] Comets E, Brendel K, Mentré F. Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: the npde add-on package for R. Comput Methods Programs Biomed. 2008;90(2):154–66.

[27] MONOLIX. MOdèles NOn LInéaires à effets miXtes). Antony, France; 2016. Available from: http://lixoft.com/products/monolix/.

[28] Desmée S, Mentré F, Veyrat-Follet C, Guedj J. Nonlinear mixed-effect models for prostate-specific antigen kinetics and link with survival in the context of metastatic prostate cancer: a comparison by simulation of two-stage and joint approaches. AAPS J. 2015;17(3):691–9.

[29] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015;65(1):5–29.

[30] Roach M, Hanks G, Thames H, Schellhammer P, Shipley WU, Sokol GH, et al. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO Phoenix Consensus Conference. Int J Radiat Oncol Biol Phys. 2006;65(4):965–74.

[31] Stephenson AJ, Kattan MW, Eastham JA, Dotan ZA, Bianco FJ, Lilja H, et al. Defining biochemical recurrence of prostate cancer after radical prostatectomy: a proposal for a standardized definition. J Clin Oncol. 2006;24(24):3973–8.

[32] Tannock IF, Fizazi K, Ivanov S, Karlsson CT, Fléchon A, Skoneczna I, et al. Aflibercept versus placebo in combination with docetaxel and prednisone for treatment of men with metastatic castration-resistant prostate cancer (VENICE): a phase 3, double-blind randomised trial. Lancet Oncol. 2013;14(8):760–8.

[33] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2017. Available from: https://www.R-project.org/.

[34] Lawrence Gould A, Boye ME, Crowther MJ, Ibrahim JG, Quartey G, Micallef S, et al. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. Stat Med. 2015;34(14):2181–95.

[35] Desmée S, Mentré F, Veyrat-Follet C, Sébastien B, Guedj J. Using the SAEM algorithm for mechanistic joint models characterizing the relationship between nonlinear PSA kinetics and survival in prostate cancer patients. Biometrics. 2017;73:305–12.

[36] Cox DR, Snell EJ. A general definition of residuals. J R Stat Soc Series B Stat Methodol. 1968;30(2):248–275.

[37] Collett D. Modelling survival data in medical research. Chapman and Hall/CRC; 2014.

[38] Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. The Annals of Statistics. 1982;10(4):1100–1120.

[39] Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. Biometrika. 1990;77(1):147–60.

[40] Hutmacher M. A Visual Predictive Check for the evaluation of the hazard function in time-to-event analyses. PAGE 22. 2013;Abstr 2940. Available from: www.page-meeting.org/?abstract=2940.

[41] Huh Y, Hutmacher M. Application of a hazard-based visual predictive check to evaluate parametric hazard models. J Pharmacokinet Pharmacodyn. 2016;43(1):57–71.

[42] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–526.

Legend to Figures

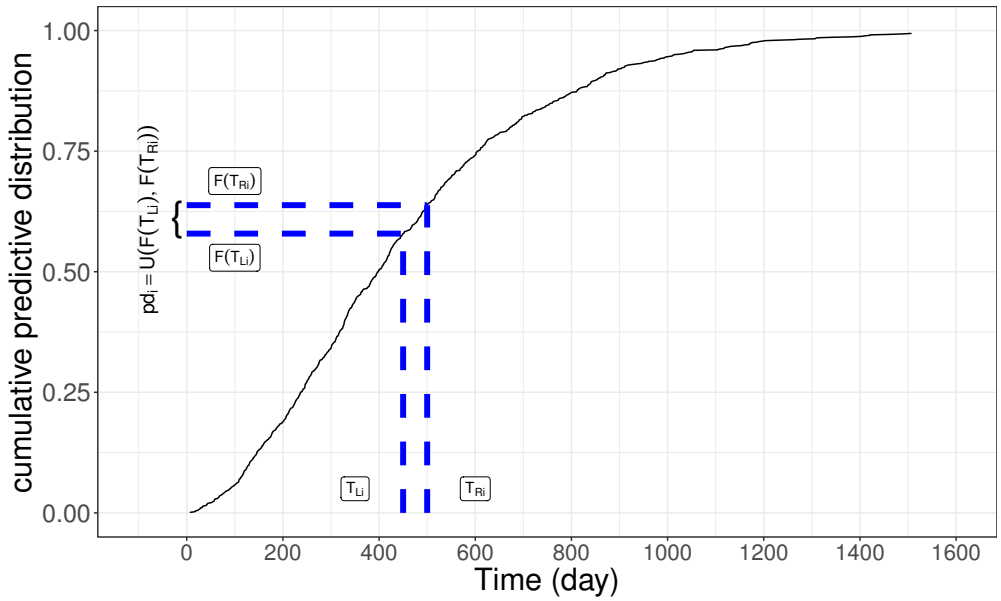

Figure 1: Example of the computation of prediction discrepancies for time-to-event observation. The black line is the cumulative predictive distribution F for one observation. T_i lies within an interval $[T_{L_i}, T_{R_i}]$ and the corresponding pd_i is sampled in a uniform distribution within the interval $[F(T_{L_i}), F(T_{R_i})]$.

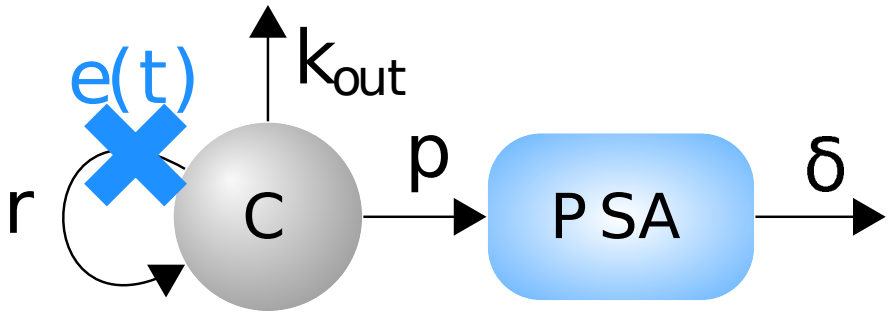

Figure 2: PSA evolution model. PSA is expressed in $ng.mL^{-1}$ and prostate cells C in mL^{-1}; r is the rate of prostate cell proliferation in the absence of treatment (day^{-1}); k_{out} the rate of prostate cell elimination (day^{-1}); p the rate of PSA secretion by C ($ng.day^{-1}$); δ the rate of PSA elimination (day^{-1}); and $e(t)$ the time-dependent treatment effect.

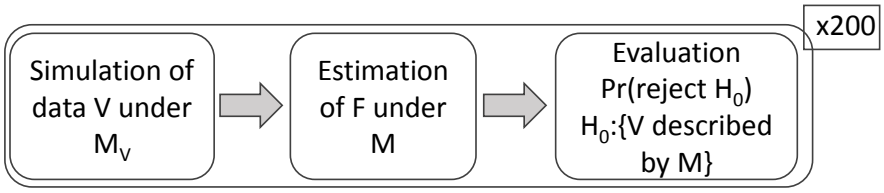

Figure 3: Schematic representation of the evaluation process. For each replication, a dataset V is simulated under model M_V. The *npde* are obtained using the cumulative predictive distribution F of an observation obtained under a model M, and we test the hypothesis that M describes the data in V. The process is repeated 200 times for each scenario.

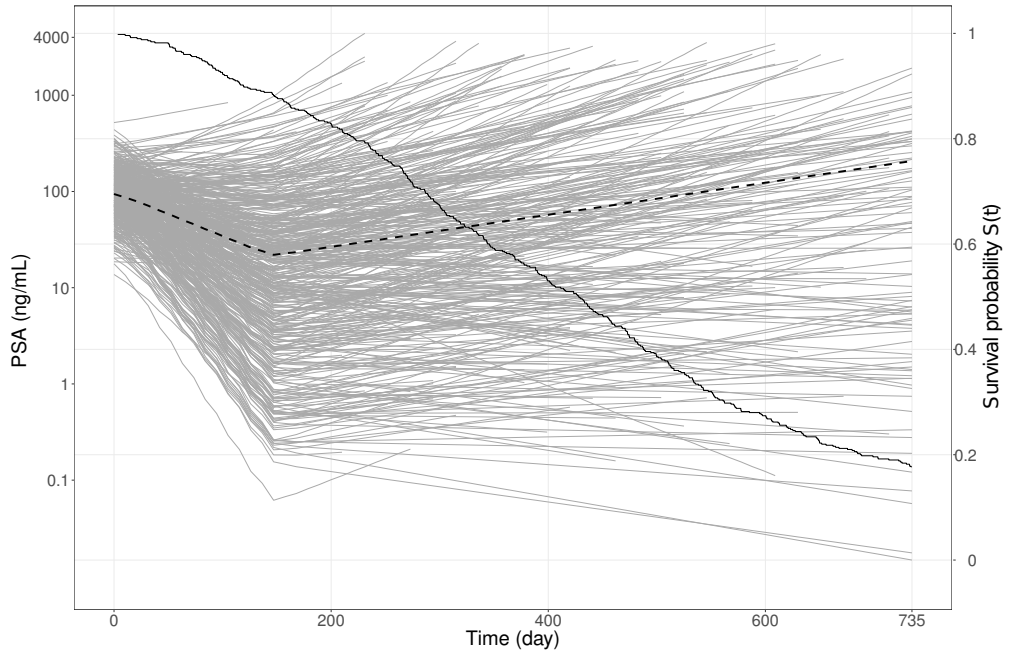

Figure 4: Spaghetti plot of the predicted PSA in one simulated dataset (grey). The PSA profile for the typical individual is shown as a dashed black line. The survival Kaplan Meier estimate for the same dataset is shown as a solid black line.

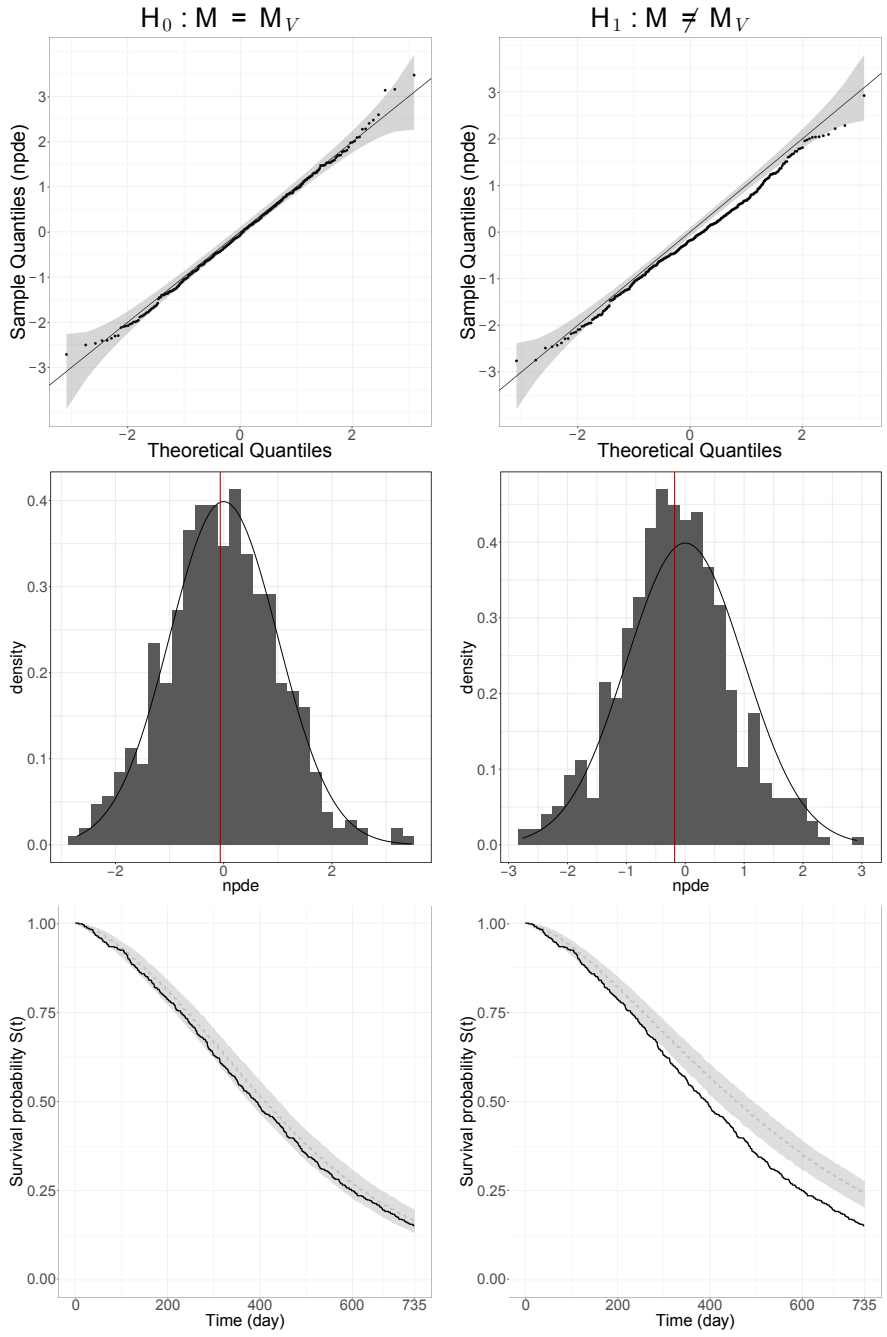

Figure 5: Top: Quantile-quantile plots of *npde* based on observations versus the theoretical standard normal distribution $\mathcal{N}(0,1)$. The theoretical distribution is represented by line $y = x$ with the 95% confidence interval. Middle: histograms of *npde* are compared to the normal density function (black curve). The median of *npde* is represented by the red vertical line. Bottom: Kaplan Meier VPC: Kaplan Meier survival estimate curve (black line) is compared to the 90% prediction interval (shaded grey area and median in dashed grey line). For graphs, the dataset was generated with $\mathcal{M}_{\mathcal{V}} = \mathcal{M}_{medium}$. In the left (resp. right) panel, the tested model $\mathcal{M}$ is $\mathcal{M}_{medium}$ (resp. $\mathcal{M}_{high}$).

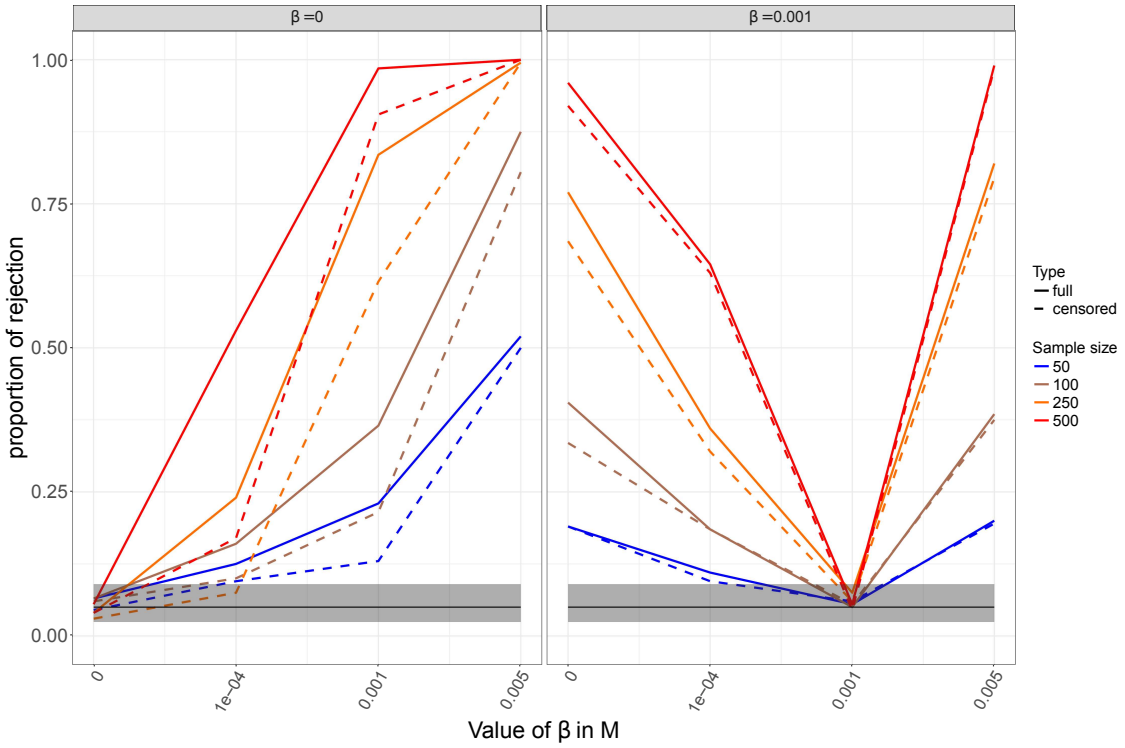

Figure 6: Performance of *npde* in two scenarios. The left-hand plot corresponds to datasets V simulated under $\mathcal{M}_{V,noLink}$ ($\beta = 0$), whereas the right-hand plot was obtained with datasets V simulated under $\mathcal{M}_{V,medium}$ ($\beta = 0.001$). The tested models are ordered on the X-axis with increasing values of β, from $\mathcal{M}_{noLink}$ ($\beta = 0$) to $\mathcal{M}_{high}$ ($\beta = 0.005$), representing an increasing strength for the link between the longitudinal biomarker and survival. When the tested model is the true model used to generate V (leftmost point in the curve when $M_V = \mathcal{M}_{V,noLink}$, and second to last point in the curve when $M_V = \mathcal{M}_{V,medium}$), we are under H_0 and the proportion of rejection corresponds to the type I error. The shaded grey area represents the prediction interval of a theoretical type I error equal to 0.05, and we expect the type I error to remain within this interval. All other cases are under H_1 and the proportion of rejection represents the power. Each colour represents the performance (type I error or power) with a different sample size, ranging from $N = 50$ to $N = 500$. Solid (resp. dashed) lines show the proportion of rejected datasets without (resp. with) censoring.

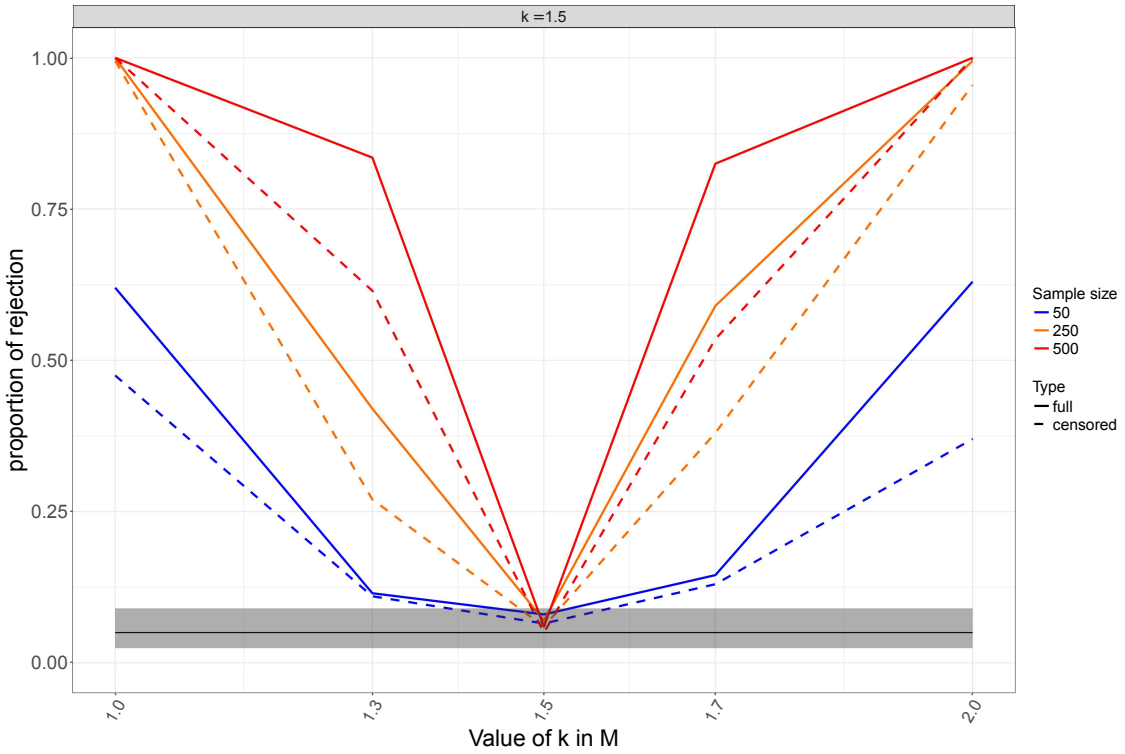

Figure 7: Type I error and power for different values of k. Each line represents the evolution of the proportion of rejection versus k for different values of N (in {50,250,500}), with solid lines for uncensored data and dashed lines in the presence of censoring.