Joint modeling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the SAEM algorithm

Cyprien Mbogning, Kevin Bleakley, Marc Lavielle

Joint modeling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the SAEM algorithm

Cyprien Mbogning[1,2,3], Kevin Bleakley[1,2] and Marc Lavielle[1,2,*]

[1]*Inria Saclay, POPIX team.*
[2]*Laboratoire de Mathématiques d'Orsay (LMO), Bat 425, 91405 Orsay cedex, France.* [3]*LIMSS, Ecole Nationale Supérieure Polytechnique (ENSP), 8390 Yaoundé, Cameroun.*
* Corresponding email: marc.lavielle@math.u-psud.fr

Abstract

We propose a nonlinear mixed-effects framework to jointly model longitudinal and repeated time-to-event data. A parametric nonlinear mixed-effects model is used for the longitudinal observations and a parametric mixed-effects hazard model for repeated event times. We show the importance for parameter estimation of properly calculating the conditional density of the observations (given the individual parameters) in the presence of interval and/or right censoring. Parameters are estimated by maximizing the exact joint likelihood with the Stochastic Approximation Expectation-Maximization algorithm. This workflow for joint models is now implemented in the Monolix software, and illustrated here on five simulated and two real data sets.

Key words: Joint models; Mixed-effects models; Repeated time-to-events; Maximum likelihood; SAEM algorithm.

1 Introduction

Joint models are a class of statistical methods for bringing together longitudinal data and time-to-event data into a unified framework. In the medical setting (the most common application of joint models), we often have, for

a set of patients, time-to-event data of interest, e.g. tumor recurrences, epileptic seizures, asthma attacks, migraines, infectious episodes, heart attacks, injuries, hospital admissions, or even death. One may be interested in modeling the process inducing the event(s), using for example a suitable chosen hazard function to describe the instantaneous chance of an event occurrence.

Simultaneously, for each patient we may be able to measure a longitudinal outcome (*biomarker* in the following) and model its progression. Joint models come into the picture when there is a distinct possibility that a given longitudinal biomarker has a real influence on the time-to-event process. In such cases and in the most general way possible, the *joint model strategy* is to suggest a relationship between the biomarker and the hazard function, i.e., have its predicted value influence the instantaneous probability of the event of interest.

Early attempts to create joint models and apply them to biological settings were introduced in [28] and [6] with applications in AIDS research. What goes today as the standard joint model was introduced in [11] and [34] and since that time, developments in the field have continued apace. We now briefly present joint modeling, then explain the contribution of the present article to the state of the art. For a more thorough introduction, we point the reader to the book [24].

Joint modeling tries to characterize the relationship between a longitudinal biomarker's evolution and the risk of a given event, while also providing an acceptable model of the biomarker's evolution itself. First, let us concentrate on the longitudinal biomarker. Its evolution is often modeled under a *linear mixed-effects* framework [17, 12, 32] using for instance splines [26] or B-splines with random effects [25, 2]. This framework takes into account the correlated nature of the measures for a given individual, while also allowing inter-individual random variability in key model parameters (e.g. slope, intercept). We can thus estimate the mean values of these parameters, as well as model/plot the evolution of the biomarker for each individual using their own estimated parameter values. Parameter estimation is often performed using a maximum likelihood strategy. However, linearity and the associated supposition of normally distributed parameters are strong hypotheses which are not necessarily representative of what is seen in real-life situations. For instance, in pharmacometrics and in particular pharmacokinetic-pharmacodynamic (PKPD) applications, linear models are usually not sufficient to satisfactorily model data. Consequently, nonlinear mixed-effects models have been largely adopted [29, 19, 5, 33] even though they involve computationally taxing calculations when performing maximum

likelihood, a stumbling block until recently. However, strategies such as the Stochastic Approximation EM (SAEM) algorithm [16], implemented in the MONOLIX software and R [23], have recently led to significantly faster methods for not only linear mixed-effects but also nonlinear mixed-effects models.

Next let us consider the event risk itself, modeled by a hazard function λ, which characterizes the distribution of the time-to-event process. The hazard function may be constant or vary as a function of time. It may or may not depend on various known or unknown population or individual variables. For instance, in the *frailty* framework [21, 20], a random multiplicative effect called "frailty" is included in the hazard function, with unknown mean and variance (to be estimated) across the population. Essentially, a more "frail" individual will have a larger multiplicative effect, and thus higher frequency of the event in question (recurrence, hospitalization, etc.). More generally, *joint modeling* is achieved by also allowing the hazard function at time t to potentially depend on the value of the longitudinal biomarker variable predicted at t. Joint modeling then involves the simultaneous estimation of all the parameters from both parts of the model. Note that under a general mixed-effects framework, one or several random effects variables can enter the longitudinal – and thus time-to-events models – in many ways, not necessarily only multiplicatively as in frailty.

Due to significant complexity in the calculation of likelihoods for joint models, initial approaches to fit them focused on two-stage methods [28, 31], with the downside of often producing biased results [4] in simulation studies. Full likelihood approaches have therefore been introduced to try to eliminate this bias [34, 13, 14]. Maximization of the log-likelihood function is then often attempted using the EM algorithm [9], treating random effects as missing data. In joint modeling using frailty, [21] and [20] use Gaussian quadrature for parameter estimation. However, Gaussian quadrature is practical only when there are a small number of random effects to be estimated. The R package **JM** [24] provides a set of procedures for solving such problems in the linear mixed-models framework.

We will show in this article that the SAEM algorithm [16] can be extended to quickly and efficiently perform joint modeling and parameter estimation in the general nonlinear framework and in the presence of censored data. To give an idea of what this means, one of our examples involves repeated events, censored data and a nonlinear continuous biomarker defined by ordinary differential equations that requires estimation of 15 parameters including 6 random effects variances. It can be solved in a few seconds.

The current article advances the state of the art in several ways. First, it presents time-to-events models for repeated events in the presence of cen-

soring under a general *nonlinear* mixed-effects framework. Second, it develops a framework for joint models combining nonlinear mixed-effects models for continuous covariates/biomarkers with (perhaps repeated) time-to-events data. Third, for likelihood calculations it presents a rigorous calculation of the conditional density of the observations given the individual parameters in a wide variety of situations (right and/or interval censored, single or multiple events). Fifth, it shows that the Stochastic Approximation Expectation Maximization (SAEM) algorithm [16] is not only capable, but also extremely fast, when it comes to performing maximum likelihood estimation for joint models. And last, it shows that we can also estimate the Fisher information matrix, the observed likelihood and the individual parameters under the same framework.

We have performed several numerical experiments to illustrate properties of the proposed methods. The experiments show that bias is introduced if we make the approximation of replacing a censoring interval by its center, or do not take into account when it is known that there is a maximum number of events. As mentioned earlier, one experiment is particularly sophisticated and requires the estimation of 15 parameters including 6 random effects variances; it takes only a few minutes to run. We then illustrate the use of these modeling methods in two real data examples: patient survival in primary biliary cirrhosis and repeated epileptic seizure count data from a clinical trial.

2 Models

2.1 Nonlinear mixed-effects models for the population approach

Consider first a single subject i of the population. Let $y_i = (y_{ij}, 1 \leq j \leq n_i)$ be the vector of observations for this subject. The model that describes the observations y_i is assumed to be a parametric probabilistic model: let $p(y_i|\psi_i)$ be the probability distribution of y_i, where ψ_i is a vector of parameters.

In the population framework, the vector of parameters ψ_i is assumed to be drawn from a population distribution $p(\psi_i;\theta)$. Then, the probabilistic model is the joint probability distribution

$$p(y_i, \psi_i; \theta) = p(y_i|\psi_i)p(\psi_i;\theta). \tag{1}$$

To define a model for the data thus consists in defining precisely these two terms.

First, let us present ψ_i in its most general form: $\psi_i = H(\psi_{pop}, \beta, c_i, \eta_i)$, where ψ_{pop} is a "typical" value of the parameters in the population, β a set of coefficients (usually called fixed effects), c_i a vector of individual covariates and η_i the random component (usually called random effects). For example, in a linear model we assume that, up to some transformation, ψ_i is a linear function of the covariates and the normally distributed random effects:

$$h(\psi_i) = h(\psi_{pop}) + \beta c_i + \eta_i, \tag{2}$$

where h is some monotonic function (log, logit, probit, etc.) and $\eta_i \sim \mathcal{N}(0, \Omega)$. The set of population parameters that define the population distribution $p(\psi_i; \theta)$ of the individual parameters ψ_i is thus $\theta = (\psi_{pop}, \beta, \Omega)$.

The conditional distribution $p(y_i|\psi_i)$ of the observations depends on the type of observations (continuous, categorical, count, time-to-event, etc.). We consider here two situations:

- observations are time-to-events, perhaps repeated (several events per individual are observed) and perhaps interval or right censored (times of events are not precisely known).
- observations are a combination of continuous values (some biomarker) and time-to-events. They are thus characterized by a joint model which describes the relationship between the two types of data.

2.2 Repeated time-to-event model

In summarizing time-to-event data, there are two main functions of interest, namely the survival function and the hazard function. The actual event time t can be regarded as the value taken by a non-negative random variable T. For the case of a single event process, the survival function $S(t)$ is defined as $S(t) = \mathbb{P}(T \geq t) = e^{-\int_0^t \lambda(u)du}$, where λ is the hazard function. In the case of a repeated events process we have instead a sequence of event times (T_j) and are now interested in the probability of an event after t_j given the previous event at t_{j-1}:

$$\mathbb{P}(T_j > t_j | T_{j-1} = t_{j-1}) = e^{-\int_{t_{j-1}}^{t_j} \lambda(u)du}.$$

Under a population framework, we suppose a parametric hazard function λ_i for each individual i: $\lambda_i(t) = \lambda(\psi_i, t)$. As an example, consider the model with constant hazard [15] given by $\lambda_i(t) = \lambda_i$. Then, the duration between successive events has an exponential distribution with parameter λ_i, and the

number of events in any interval of length Δ has a Poisson distribution with parameter $\Delta\lambda_i$. Here, the vector of individual parameters reduces to $\psi_i = \lambda_i$.

In the most simple case, y_i is a vector of known event times: $y_i = (t_{i1}, t_{i2}, \ldots, t_{i\,n_i})$. But if we only know that events occur within certain intervals, then observations are the *number of events per interval*. Let $(I_{i1}, \ldots, I_{i\,n_i})$ be a set of disjoint time intervals for individual i relevant to the experimental design. We then can write $y_i = (k_{i1}, \ldots, k_{i\,n_i})$, where $k_{i\ell}$ is the number of events for individual i that have occurred in interval $I_{i\ell}$. Note that this includes the interval censored case with finite intervals $I_{i\ell}$, as well as the right censored case with $I_{i\,n_i} = [t_{\text{end}}\,, \infty)$.

2.3 Joint models

Besides the parametric form of the model, an essential point of joint modeling is the type of dependency between the longitudinal data model and the events. Suppose that we have a continuous biomarker of the form

$$b_{ij} = f\left(t_{ij}, \psi_i^{(1)}\right) + g\left(t_{ij}, \psi_i^{(1)}\right)\varepsilon_{ij}, \qquad 1 \leq i \leq N,\ 1 \leq j \leq n_{1,i}, \tag{3}$$

where ε_{ij} is a residual error with mean 0 and variance 1. Next, we connect this with RTTE via the hazard function given in general form: $\lambda_i(t) = \lambda\left(f(t, \psi_i^{(1)}), \psi_i^{(2)}\right)$. Observations are therefore a combination of the $n_{1,i}$ continuous-valued biomarker measurements with the $n_{2,i}$ event times (if observed): $y_i = ((b_{ij}, 1 \leq j \leq n_{1,i}), (t_{i\ell}, 1 \leq \ell \leq n_{2,i}))$, or with the number of events per interval in the case of censoring: $y_i = ((b_{ij}, 1 \leq j \leq n_{1,i}), (k_{i\ell}, 1 \leq \ell \leq n_{2,i}))$. The vector of individual parameters $\psi_i = (\psi_i^{(1)}, \psi_i^{(2)})$ combines the individual parameters from the two parts of the joint model.

Example. Suppose that the biomarker measurements can be modeled by

$$b_{ij} = \gamma_i + \delta_i t_{ij} + a_i\varepsilon_{ij}, \qquad 1 \leq i \leq N,\ 1 \leq j \leq n_{1,i} \tag{4}$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$, and that they are related to an event process with hazard function

$$\lambda_i(t) = \lambda_{0,i} e^{\alpha_i(\gamma_i + \delta_i t)}. \tag{5}$$

Here, $\psi_i = (\gamma_i, \delta_i, a_i, \lambda_{0,i}, \alpha_i)$. Formulas 4 and 5 thus characterize the probability model $p(y_i|\psi_i)$ of the observations. In the following, we will suppose certain parameters to be constants (i.e., without inter-individual variability),

essentially for reasons of identifiability. This is not to do with the distribution of the observations, but rather with the distributions of the individual parameters. In effect, building the statistical model for the individual parameters relies in part on deciding which components of ψ_i vary or not within the population.

3 Tasks and methods

There are a variety of tasks that we are interested in performing here, whether it be estimating population parameters and their variation, estimating individual parameters or estimating the likelihood, the latter useful for performing likelihood ratio tests and calculating information criteria such as BIC.

In the following sections, we propose methodology for each of these tasks. We show that each requires calculation of the joint pdf (1) and in particular the conditional density $p(y_i|\psi_i)$ as under (2), $p(\psi_i;\theta)$ is straightforward to compute since it is derived from a Gaussian density.

3.1 Maximum likelihood estimation of the population parameters

Estimation in mixed-effects models consists of estimating the probability distribution of the ψ_i's in the population from the observations of the N subjects, i.e., in evaluating both the typical values in the population and the variability between subjects. More precisely, we aim to compute the maximum likelihood estimate of θ in (nonlinear) mixed-effects models by maximizing the observed likelihood $p(y;\theta)$. Estimation is complex because the N random vectors of parameters ψ_i are not observed and because there is a nonlinear relashionship between the observations and the random effects defined in (2). For these reasons the likelihood function can not be explicitly given and its maximization is far from straightforward.

In a general way, linear and nonlinear mixed-effects models, including mixed-effects diffusion models, can be seen as incomplete data models in which the individual parameters $\pi = (\psi_1, \ldots, \psi_N)$ are the non-observed data and the population parameters are the parameters of the model that need to be estimated from the N individual observations vectors $y = (y_1, \ldots, y_N)$. The EM algorithm ([9]) iteratively performs parameter estimation in such models. The algorithm requires computing at each iteration the conditional expectation $E\left(\log p(y,\psi;\theta)|y,\theta^{(k-1)}\right)$, where $\theta^{(k-1)}$ represents the current

estimation of θ. In many situations, especially when dealing with nonlinear mixed-effects models, this conditional expectation has no closed form. Variants of the algorithm get around this difficulty.

For instance, in the SAEM algorithm [8], the E-step is evaluated by a stochastic approximation procedure. In Web Appendix A, a detailed description of the SAEM algorithm is presented. SAEM is extremely fast and has already been used to help treat a large range of real-world problems including Hepatitis C treatment outcomes [30], longitudinal data analysis [27], parameter estimation in HIV models [3, 18], bioequivalence crossover trials [10], and much more. The MONOLIX software provides a general implementation of SAEM that can be easily extended by the user to new modeling challenges. SAEM has also been implemented in the R package `saemix` and the Matlab statistics toolbox as `nlmefitsa.m`.

It turns out that in this framework, computation of $p(y_i|\psi_i)$ for the various cases (repeated events with interval censoring, right-censored time-to-events, joint models, etc.) is a critical modeling step. In the following section, we therefore explicitly calculate this pdf for a wide range of cases.

3.2 Computing the probability distribution for repeated time-to-events

The aim of this section is to provide the precise expression of the conditional distribution $p(y_i|\psi_i)$ for any subject i, when the vector of observations y_i only consists of (possibly repeated and possibly censored) time-to-events. While some of these results are known, it is useful to restate them here in a "ready-to-use" form. For a much more expansive treatment, we refer the reader to the monograph [1]. For the sake of simplicity we only consider a single subject, and therefore omit the subscript i in notation. Also for simplicity we denote $\lambda(t)$ the hazard function at time t and omit the dependence with respect to the parameter ψ. We assume that the trial starts at time t_0 and ends at time t_{end}. Both t_0 and t_{end} are known. Let $T = (T_1, T_2, \ldots)$ be the (random) event times after t_0, and Λ be the cumulative hazard function:

$$\Lambda(a,\ell) \quad := \quad \int_a^\ell \lambda(t)\,dt \quad = \quad \Lambda(t_0,\ell) - \Lambda(t_0,a).$$

By definition, recall that

$$\mathbb{P}(T_j > t_j | T_{j-1} = t_{j-1}) = e^{-\Lambda(t_{j-1},t_j)}. \tag{6}$$

In the following, we distinguish between exactly observed and interval-censored events. In each case, we further distinguish between whether the last event

occurred before the end of the experiment or if it is right censored, as the value of the conditional distribution turns out to be different for each. For conciseness, explicit derivations of the following results have been placed in Web Appendix B.

Let $k_{\max}$ ($k_{\max} \leq +\infty$) be the maximum number of events that can occur. $k_{\max}$ can be either bounded ($k_{\max} = 1$ for events such as death) or unbounded ($k_{\max} = +\infty$ for seizures, hemorrhaging ,...).

3.2.1 Exactly observed events

i) the last event is observed. Assume that we observe $n = k_{\max}$ events at times t_1, t_2, ..., t_n. Here, $n = k_{\max}$ means that no event will occur after t_{end}. The vector of observations is $y = (t_1, t_2, \ldots, t_n)$, and

$$p(y|\psi) \quad = \quad \prod_{j=1}^{n} \lambda(t_j) e^{-\Lambda(t_{j-1}, t_j)} \,. \tag{7}$$

ii) the last event is not observed. Assume that we observe $n < k_{\max}$ events at times t_1, t_2, ..., t_n. $n < k_{\max}$ means that an event will to occur at a certain unknown time $T_{n+1} > t_{end}$. Here, the vector of observations is $y = (t_1, t_2, \ldots, t_n, t_{n+1} > t_{end})$, and

$$p(y|\psi) \quad = \quad \left(\prod_{j=1}^{n} \lambda(t_j) e^{-\Lambda(t_{j-1}, t_j)} \right) e^{-\Lambda(t_n, t_{end})} \,.$$

3.2.2 Interval censored events

Consider first a single interval $[0, \ell]$ and let $k_{\max}$ ($k_{\max} \leq \infty$) be the maximum number of events that can occur. $k_{\max}$ can either be bounded ($k_{\max} = 1$ for events such as death) or unbounded ($k_{\max} = \infty$ for seizures, hemorrhaging, etc.). Let K be the number of events in $[0, \ell]$. For any $k < k_{\max}$, $K = k$ implies that the $(k+1)$-th event occurs after time ℓ. Then, for any $k < k_{\max}$, it is well known that

$$\mathbb{P}(K = k) \quad = \frac{\Lambda(0, \ell)^k}{k!} e^{-\Lambda(0, \ell)}. \tag{8}$$

So, for a bounded number of events ($k_{\max} < +\infty$),

$$\mathbb{P}(K = k_{\max}) \quad = \quad 1 - \sum_{k=0}^{k_{\max}-1} \mathbb{P}(K = k) \quad = \quad 1 - \sum_{k=0}^{k_{\max}-1} \frac{\Lambda(0,\ell)^k}{k!} e^{-\Lambda(0,\ell)}. \tag{9}$$

Consider now n contiguous intervals ($[\ell_{j-1}, \ell_j]; 1 \leq j \leq n$), where $\ell_0 = t_0$ and $\ell_n = t_{end}$. Let K_j be the number of events in interval $[\ell_{j-1}, \ell_j]$.

i) the last event is observed.. Let $s_{n-1} = \sum_{j=1}^{n-1} k_j$. Using equations 8 and 9, we can show that

$$p(y|\psi) \quad = \quad \left(\prod_{j=1}^{n-1} \frac{\Lambda(\ell_{j-1}, \ell_j)^{k_j}}{k_j!} e^{-\Lambda(\ell_{j-1},\ell_j)} \right) \tag{10}$$

$$\times \left(1 - \sum_{k=0}^{k_{\max}-s_{n-1}} \frac{\Lambda(\ell_{n-1}, \ell_n)^k}{k!} e^{-\Lambda(\ell_{n-1},\ell_n)} \right). \tag{11}$$

ii) the last event is not observed. This implies that the first non-observed event occurs after t_{end}. Using equation (11), it is straightforward to show that if $\sum_{j=1}^{n} k_j < k_{\max}$, then

$$p(y|\psi) \quad = \quad \prod_{j=1}^{n} \left(\frac{\Lambda(\ell_{j-1}, \ell_j)^{k_j}}{k_j!} e^{-\Lambda(\ell_{j-1},\ell_j)} \right). \tag{12}$$

4 Simulations and applications

4.1 Simulations

A series of simulation studies were conducted to evaluate the proposed methodology for calculating the maximum likelihood estimate of the population parameters. The first three consider only time-to-events in order to illustrate the statistical properties of the maximum likelihood estimator and to show why censoring needs to be correctly taken into account. The fourth trial presents joint modeling for the example given in Section 2.3. Due to space requirements, a further example of sophisticated joint modeling of a pharmacokinetics problem is left to Web Appendix C. It combines almost everything we can throw at it: repeated events, censored data, and a nonlinear continuous biomarker defined by ordinary differential equations that itself

can be censored ("below limit of quantification"). The model requires estimation of 15 parameters including 6 random effects variances, yet takes only a few second to converge to systematically accurate parameter estimates. In Web Appendix C, we also briefly present model diagnotic tools, even though this is beyond the scope of the present paper.

For each scenario, the SAEM algorithm was used with $M = 100$ simulated datasets for computing the parameter estimates $(\hat{\theta}_m, 1 \leq m \leq M)$. To assess statistical properties of the proposed estimators for each parameter, percentage-wise relative estimation errors $(REE_m, 1 \leq m \leq M)$ were computed:

$$REE_m = \frac{\hat{\theta}_m - \theta^\star}{\mid \theta^\star \mid} \times 100.$$

Using the REEs, the relative bias (RB) and relative root mean square errors (RRMSE) were computed for each parameter in each scenario:

$$\begin{aligned} RB &= \frac{1}{M} \sum_{m=1}^{M} REE_m \\ RRMSE &= \sqrt{\frac{1}{M} \sum_{m=1}^{M} REE_m^2}\,. \end{aligned}$$

Also, for each scenario the Fisher information matrix was estimated and standard errors $(\widehat{se}_m, 1 \leq m \leq M)$ of the estimated parameters derived. Of course, the true standard errors $se^\star$ are unknown, but they can be empirically estimated by the root mean square errors (RMSE) of the estimated parameters:

$$RMSE = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{\theta}_m - \theta^\star)^2}.$$

To assess statistical properties of the proposed estimator of the standard errors, we can then compare them with the RMSE by computing relative estimation errors (in %) for each replicate:

$$REEse_m = \frac{\hat{se}_m - RMSE}{\mid \theta^\star \mid} \times 100.$$

4.1.1 Example 1

This first example demonstrates that an accurate estimation of the interpatient variability of the hazard function requires observation of multiple

events. We consider a basic RTTE model under the mixed-effects framework with constant hazard (See [15]) for each individual $i = 1, 2, \dots N$, expressed as

$$\lambda_i(t) = \lambda_i \tag{13}$$
$$\log(\lambda_i) \sim \mathcal{N}\left(\log(\lambda), \omega^2\right). \tag{14}$$

The goal in mixed-effects modeling is then to estimate the population value λ and variance ω^2. We assume that events are observed between time $t_0 = 0$ and time $t_{\text{end}} = 12$. Furthermore the event times are assumed to be exactly known. We are thus in the situation described Section 3.2.1 with observed and right censored events. The conditional distribution of the observations is given in equation 7. The 100 datasets with 120 individuals in each were simulated under nine different scenarios with $\lambda \in \{0.01,\ 0.1,\ 1\}$ and $\omega \in \{0.1,\ 0.5,\ 1\}$. The distributions of the REE_m and $REEse_m$ are displayed in Figures 1 and 2.

Figures 1 and 2 display the relative estimation errors for λ and ω and for their respective standard errors, obtained with 9 different scenarios ($\lambda = 0.01, 0.1, 1$ and $\omega = 0.1, 0.5, 1$). This figures show that λ and its standard error are well-estimated generally and that the estimator is essentially unbiased. We see also that ω and its standard error are poorly estimated when both λ and ω are small, but as the true value of λ increases (i.e., more events happen), estimation of ω significantly improves and becomes unbiased.

Figure 3 shows a 2-d log-likelihood profile for the parameter combinations $(\lambda^\star, \omega^\star) = (0.01, 0.5)$ on the left and $(\lambda^\star, \omega^\star) = (0.1, 0.5)$ on the right, and shows the disparity between the true parameters and the maximum likelihood estimates from two simulation runs. We see that the log-likelihood is much more concentrated around its global maximum when $\lambda = 0.1$, and maximization of the log-likelihood cannot provide an accurate estimation of ω if $\lambda = 0.01$, i.e. if the number of events is too small.

Note that Gaussian quadrature, efficient for low-dimensional numerical integration, was used to compute the log-likelihood colormap.

4.1.2 Example 2

The goal of this experiment is to shown that interval censoring should be properly taken into account in order to avoid estimation bias. Using the same basic model as the previous example, the simulation scheme is the following:

- $M = 100$ datasets with $N = 1000$ subjects in each.

- The true parameter values are $(\lambda^\star, \omega^\star) = (0.5, 0.5)$.
- The event process is single-event (e.g., death).
- Events are interval or right-censored. The intervals are contiguous and of length Δ.
- Observations occur between $t_0 = 0$ and $t_{\text{end}} = 24$.

Here, the events are interval or right-censored. If we incorrectly take this information into account, for instance by considering that the event has happened at the interval midpoint, decreasing estimation quality and bias are introduced as the length Δ of the intervals increases. Figure 4 shows that very little information is lost (with respect to the case where we *do* know the exact times) if the correct formula is applied, whereas if the model is misspecified, the RRMSE and relative bias increase considerably as Δ increases.

4.1.3 Example 3

The goal of this experiment is to show that the maximum number of possible events is a piece of information that needs to be taken into account in order to avoid estimation bias. Here, we take the same interval or right-censored model as the previous examples but this time, we suppose that there are a maximum of $k_{\max} = 5$ events per subject. Note that the $k_{\max} = 5$ events are not necessarily observed during the trial period.

If for subject i we *have* observed $k_{\max} = 5$ events, the correct formula is (11), because it takes into account the fact that there are a maximum of 5 events and they have all been observed. If on the other hand the last event or events have not been observed, then equation (12) should be used when performing maximum likelihood estimation. Figure 5 shows what happens when this is not correctly taken into account.

Indeed, as the width Δ of the intervals increases, the RRMSE and the (absolute) relative bias increase markedly in the misspecified case with respect to the correct one.

4.1.4 Example 4

The goal of this experiment is to show that SAEM performs very well for estimating the parameters of a joint model, event when the events are interval censored and when the continuous data model is nonlinear. The example, first introduced in Section 2.3, can be seen as an extension of the previous

ones to joint modeling, or analogously as an RTTE model with a time-dependent covariate, taken here as a biomarker. We consider thus a joint model with a biomarker representing disease progress and an event which can occur several times during the study. The model is:

$$\begin{aligned} b_{ij} &= \gamma_i + \delta_i t_{ij} + a\varepsilon_{ij}, \qquad 1 \leq i \leq N,\ 1 \leq j \leq n_{1,i} \\ \lambda_i(t) &= \lambda e^{\alpha(\gamma_i + \delta_i t)}, \end{aligned}$$

where $\log(\gamma_i) \sim \mathcal{N}(\log \gamma_{\text{pop}} + \beta \text{C}_i, \omega_\gamma^2)$ and $\log(\delta_i) \sim \mathcal{N}(\log \delta_{\text{pop}}, \omega_\delta^2)$. We suppose that $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$.

Here, the hazard increases exponentially as the disease progresses linearly. C_i represents the treatment covariate, which takes values 0 (untreated) and 1 (treated). Treatment is associated with an effect which produces an immediate reduction of the slope of the disease progress.

Remark 1. Methods developed for linear models (e.g. [34, 13, 14]) cannot be used here since the model is not linear with respect to the random effects: the continuous observations are not normally distributed since γ_i and δ_i are not normally distributed. A more advanced method, such as SAEM for example, is therefore required for estimating the population parameters of the model. Furthermore, the exponential term of the hazard contains several individual (random) parameters, and is thus more general than frailty-based methods.

Remark 2. We do not focus on model selection in the present paper. Our goal is to provide a powerful methodology that can be implemented for a wide and general range of models chosen by the user. Tools for model diagnosis are beyond the scope of the paper, though the interested reader may refer to the simulations in Web Appendix C for an example of this.

For the model in question, we consider the following design:

- the biomarker (b_{ij}) is observed at times 0, 25 and 50 weeks: ($n_{i,1} = 3$).
- possibility of repeated interval-censored events until the end of the experiment at $t_{\text{end}} = 50$ weeks. Observations are the number of events in each 5 week period between $t_0 = 0$ and $t_{\text{end}} = 50$ (so $k_{i,1}$ is the number of events between 0 and 5,..., $k_{i,10}$ the number of events between 45 and 50). Thus, $n_{i,2} = 10$.
- parameter values are $\gamma_{\text{pop}} = 1$, $\delta_{\text{pop}} = 100$, $\beta = -0.3$, $\alpha = 0.02$, $\lambda = 0.01$, $\omega_\gamma = 0.1$, $\omega_\lambda = 0.1$, $a = 1$.

We suppose the total number of subjects in the trial is $N = 1000$. Note that we take a large N here because our fundamental goal is not to show in detail the performance of the maximum likelihood estimation. Rather, it is to show that the SAEM algorithm is effective in this framework: non-linear model, repeated interval-censored events, that it is fast and that it leads to little or no bias as well as small REEs.

SAEM performed well with the given model and experimental design. First, it was fast, taking 82 seconds on an Intel(R) Core(TM) i7-2760QM laptop with a 2.4 GHz processor. Figure 6 shows the convergence of the parameter estimates in a typical run, requiring less than 100 iterations for all parameters. Figure 7 shows that there is little or no bias in the parameter estimation, and consistently small REEs across the trials.

4.1.5 Example 5

This example from pharmacokinetics, described in detail in Web Appendix C, shows that it is possible to perform joint modeling when the model for the longitudinal variable depends on a set of differential equations and the hazard function is dependent on both time and several (random) individual parameters.

4.2 Applications

4.2.1 Primary Biliary Cirrhosis Data

This well known dataset comes from a study conducted by the Mayo Clinic from 1974 to 1984. The study includes 158 patients who received D-penicillamine and 154 who received a placebo. Patient survival is the outcome of main interest. By the end of the study, 140 patients had died and 172 were still alive. Several biomarkers, including serum bilirubin, were measured during the study. A total of 1945 measurements of serum bilirubin were made available.

Various joint models for this data are proposed in [24]. All of these assume a linear mixed-effects model for the longitudinal data. We will show that our approach provides a straightforward extension to more general non-linear mixed-effects models.

Following [24], we used the following model for the serum bilirubin: $m_i(t) = c_{0,i} + c_{1,i}t + c_{2,i}t^2$, with $\log b_{ij} = m_i(t_{ij}) + a\varepsilon_{ij}$. Here, $m_i(t)$ is the predicted concentration of bilirubin for patient i at time t and b_{ij} its measured concentration at time t_{ij}. We used a simple proportional hazard model for the survival data: $\lambda_i(t) = \lambda_{0,i}e^{\alpha_i\, m_i(t)}$.

The vector of individual parameters for patient i is given by $\psi_i = (c_{0,i}, c_{1,i}, c_{2,i}, \lambda_{0,i}, \alpha_i)$. Different statistical models for the ψ_i's were compared initially assuming Gaussian distributions for the $c_{\ell,i}$ and fixed parameters $\lambda_{0,i} = \lambda_0$ and $\alpha_i = \alpha$.

We then considered a latent class model for the longitudinal data, assuming that the population is heterogeneous and constituted of two subpopulations that cannot be clearly identified by any of the available covariates. In other words, we assumed a mixture of two Gaussian distributions for the $c_{\ell,i}$. Note again that this example is for illustrating the general methodology, and not model selection (i.e., selecting the "best" number of subpopulations), which is beyond the scope of the paper.

Let (z_i) be a sequence of latent variables such that $z_i = 0$ if patient i belongs to subpopulation 1 and $z_i = 1$ if they belong to subpopulation 2. We also introduce the treatment (D-penicillamine/placebo) as a categorical covariate: let (d_i) be a sequence of observed variables such that $d_i = 0$ if patient i receives the placebo and $d_i = 1$ if the patient receives the active treatment. The statistical model for the individual parameters can therefore be described as follows:

$$
\begin{aligned}
c_{0,i} &= c_0 + \beta_{0,z} z_i + \beta_{0,d} d_i + \eta_{0,i}, \quad \eta_{0,i} \sim \mathcal{N}(0, \omega_0^2) \\
c_{1,i} &= c_1 + \beta_{1,z} z_i + \beta_{1,d} d_i + \eta_{1,i}, \quad \eta_{1,i} \sim \mathcal{N}(0, \omega_1^2) \\
c_{2,i} &= c_2 + \beta_{2,z} z_i + \beta_{2,d} d_i + \eta_{2,i}, \quad \eta_{2,i} \sim \mathcal{N}(0, \omega_2^2) \\
\lambda_{0,i} &= \lambda_0 + \beta_{\lambda,d} d_i \\
\alpha_i &= \alpha + \beta_{\alpha,d} d_i.
\end{aligned}
$$

A diagonal variance-covariance matrix is assumed for the random effects.

Extensions of the SAEM algorithm for mixtures of mixed-effects models have been developed and implemented in MONOLIX. We combined this method for mixture models with the proposed methods for joint models in order to simultaneously fit the longitudinal and survival data. Table 1 provides parameter estimates for the model. We see that there is no significant effect of the treatment on serum bilirubin or the survival probability. Also, two different typical profiles appear to describe the serum bilirubin kinetics because the distribution of (c_0, c_2) is well characterized by a mixture of two Gaussian distributions.

4.2.2 Epileptic seizure counts

In this study, all recruited patients were on standard anti-epileptic therapy and completed a 12 week baseline screening phase. Thereafter, patients

were randomized to parallel treatment groups receiving placebo or active treatment (gabapentin 0.45, 0.6, 0.9, 1.2 and 1.8g). Overall, time profiles from 788 patients were included in the database. Data consisted of baseline daily counts of epileptic seizures measured over 12 weeks followed by 12 weeks of active treatment.

Several count data models have been proposed, including a mixture of two Poisson models [22] and a hidden Markov model [7]. Such models assume that the probability function of the number of seizures is piecewise-constant over time. We propose to extend this approach by considering seizures as interval-censored events. Then, following section 3.2.2, it is equivalent to consider the seizure count as a nonhomogenous Poisson process whose intensity is a continuous function of time. The hazard function is then modeled assuming a constant hazard in both phases and a smooth transition between the two phases:

$$\lambda_i(t) = \begin{cases} a_i & \text{if } t \leq t_0 \\ b_i + (a_i - b_i)e^{-c_i(t-t_0)} & \text{if } t > t_0, \end{cases}$$

where t_0 is the time when the active treatment starts. We used the following statistical model for describing inter-patient variability of the individual parameters a_i, b_i and c_i:

$$\begin{aligned} \log(a_i) &= \log(a) + \eta_{a,i}, \quad \eta_{a,i} \sim \mathcal{N}(0, \omega_a^2) \\ \log(b_i) &= \log(b) + \beta_b \log(1 + D_i) + \eta_{b,i}, \quad \eta_{b,i} \sim \mathcal{N}(0, \omega_b^2) \\ \log(c_i) &= \log(c) + \beta_c \log(1 + D_i), \end{aligned}$$

where D_i is the amount of gabapentin administered to patient i.

The estimated parameters are displayed Table 2 and the distributions of the hazard functions associated to different doses of gabapentin are displayed Figure 8. Even though the inter-patient variability of the hazard function is large, we can see a slight placebo effect and a mild effect of gabapentin on the seizure rate.

5 Discussion

Joint modeling of longitudinal biomarkers and time-to-events data is an important step in the improvement in understanding of the connection between biological changes in time and the arrival of a (perhaps critical) event to the patient. In recent years, linear mixed-effects models have been coupled with time-to-single event processes and parameter estimation has been performed,

often using maximum likelihood coupled with the EM algorithm. When there are more than a negligible number of random effects in the model, likelihood calculations are a huge bottleneck, discouraging use of these methods.

Here, we have shown that the SAEM algorithm is extremely capable in performing parameter estimation for joint models where the mixed-effects can be nonlinear, the events can be repeated, and all of this in the presence of right and/or interval censoring. To be able to implement SAEM for joint models in the afore-mentioned range of cases, we have for each derived precise expressions for the conditional likelihood of the observations given the individual parameters. In a series of simulation studies, we have shown that the SAEM algorithm converges for joint models in a matter of seconds or minutes rather than hours or days. As a consequence, we can also quickly estimate the Fisher information matrix, the observed likelihood and the individual parameters.

SAEM for joint models is intuitively implemented in the MONOLIX software: in order to pass from nonlinear mixed effects modeling to joint modeling, *all that is required* of the modeler is to provide the parametric form of the hazard function. We have illustrated this by performing joint modeling in two real examples: survival data and repeated time-to-event data. Note that several diagnostic tools are also implemented in MONOLIX based on Kaplan-Meier plots; further details are beyond the scope of the article.

In conclusion, now that there exists a simple, fast and high-performance tool for joint modeling, we believe these methods should now be used more in everyday statistical practice.

6 Supplementary material

Web Appendices for Sections 3.1, 3.2 and 4.1 are online at http://TBA

References

[1] O. Aalen, O. Borgan, and H. Gjessing. *Survival and Event History Analysis.* Springer, New York, 2008.

[2] E. Brown, J. Ibrahim, and V. DeGruttola. A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61:64–73, 2005.

[3] P. Chan, P. Jacqmin, M. Lavielle, L. McFadyen, and B. Weatherley. The Use of the SAEM Algorithm in MONOLIX Software for Estima-

tion of Population Pharmacokinetic-Pharmacodynamic-Viral Dynamics Parameters of Maraviroc in Asymptomatic HIV Subjects. *Journal of Pharmacokinetics and Pharmacodynamics*, 38:41–61, 2011.

[4] U. Dafni and A. Tsiatsis. Evaluating surrogate markers of clinical outcome measured with error. *Biometrics*, 54:1445–1462, 1998.

[5] M. Davidian and D. M. Giltinan. *Nonlinear Models for Repeated Measurements Data*. Chapman & Hall., London, 1995.

[6] V. DeGruttola and X. Tu. Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014, 1994.

[7] M. Delattre, R. Savic, R. Miller, M. Karlsson, and M. Lavielle. Analysis of exposure-response of CI-945 in patients with epilepsy: application of novel mixed hidden Markov modeling methodology. *J. Pharmacokinet. Pharmacodyn.*, 39:263–271, 2012.

[8] B. Delyon, M. Lavielle, and E. Moulines. Convergence Of A Stochastic Approximation Version Of The EM Algorithm. *The Annals Of Statistics*, 27:94–128, 1999.

[9] A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Roy. Stat. Soc. B*, 39:1–38, 1977.

[10] A. Dubois, M. Lavielle, S. Gsteiger, E. Pigeolet, and F. Mentré. Model-Based Analyses of Bioequivalence Crossover Trials Using the SAEM Algorithm. *Statistics in Medicine*, 30:582–600, 2011.

[11] C. Faucett and D. Thomas. Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Stat. Med.*, 15:1663–1685, 1996.

[12] D. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Biometrika*, 61:383–385, 1977.

[13] R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1:465–480, 2000.

[14] F. Hsieh, Y. Tseng, and J. Wang. Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62(1037–1043), 2006.

[15] K. E. Karlsson, E. L. Plan, and M. O. K Karlsson. Performance of three estimation methods in repeated time-to-event modeling. *The AAPS Journal*, 13(1):83–91, 2011.

[16] E. Kuhn and M. Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.*, 49:1020–1038, 2005.

[17] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[18] M. Lavielle, A. Samson, A. K. Fermin, and F. Mentré. Maximum likelihood estimation of long term HIV dynamic models and antiviral response. *Biometrics*, 67:250–259, 2011.

[19] M. J. Lindstrom and D. M. Bates. Nonlinear mixed-effects models for repeated measures. *Biometrics*, 46:673–687, 1990.

[20] L. Liu and X. Huang. Joint analysis of correlated repeated measures and recurrent events processes in the presence of a dependent terminal event. *J. ROY. STAT. SOC. C-APP.*, 58:65–81, 2009.

[21] L. Liu, X. Huang, and J. O'Quigley. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*, 64(3):950–958, 2004.

[22] R. Miller, B. Frame, B. Corrigan, P. Burger, H. Bockbrader, E. Garofalo, and R. Lalonde. Exposure response analysis of pregabalin add-on treatment of patients with refractory partial seizures. *Clin Pharmacol Ther*, 73:491–505, 2003.

[23] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[24] D. Rizopoulos. *Joint Models for Longitudinal and Time-to-Event Data. With Applications in R.* Chapman & Hall/CRC Biostatistics, Boca Raton, 2012.

[25] D. Rizopoulos, G. Verbeke, and G. Molenberghs. Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *J. Roy. Stat. Soc. B*, 71:637–654, 2009.

[26] D. Ruppert, M. Wand, and R. Carroll. *Semiparametric Regression.* Cambridge University Press, Cambridge, 2003.

[27] A. Samson, M. Lavielle, and F. Mentré. The SAEM algorithm for group comparison tests in longitudinal data analysis based on nonlinear mixed-effects model. *Statistics in Medicine*, 26:4860–4875, 2007.

[28] S. Self and Y. Pawitan. Modeling a marker of disease progression and onset of disease. In *AIDS Epidemiology: Methodological Issues.* Birkhaüser, Boston, 1992.

[29] L. B. Sheiner and S. L. Beal. Pharmacokinetic parameter estimates from several least squares procedures: superiority of extended least squares. *J. Pharmacokinet. Biop.*, 13:185–201, 1985.

[30] E. Snoeck, P. Chan, M. Lavielle, P. Jacqmin, N. Jonsson, K. Jorga, T. Goggin, S. Jumbe, and N. Frey. Hepatitis C Viral Dynamics Explaining Breakthrough, Relapse or Response after Chronic Treatment. *Clinical Pharmacology and Therapeutics*, 87(6):706–713, 2010.

[31] A. Tsiatis, V. DeGruttola, and M. Wulfsohn. Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 count in patients with AIDS. *J. Am. Stat. Assoc.*, 90:27–37, 1995.

[32] Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data.* Springer-Verlag, New York, 2000.

[33] E. G. Vonesh and V. M. Chinchilli. *Linear and nonlinear models for the analysis of repeated measurments.* Marcel Dekker, New York, 1997.

[34] M. Wulfsohn and A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339, 1997.

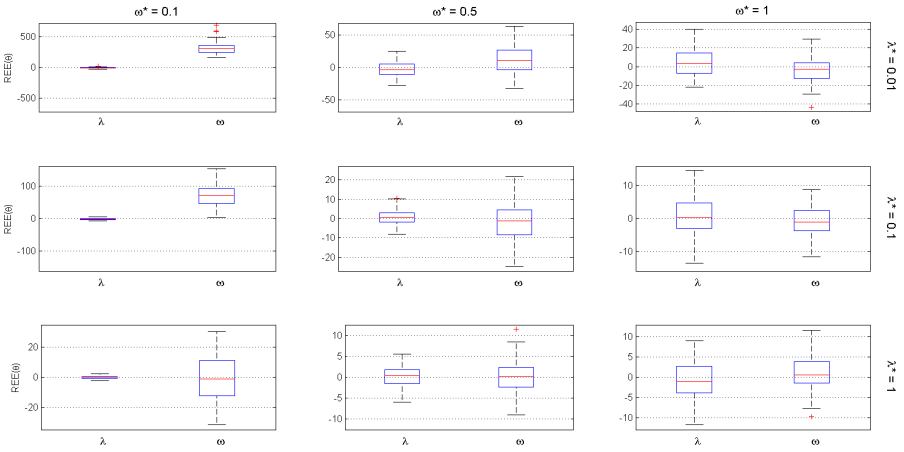

Figure 1: Relative estimation errors (in %) for λ and ω obtained with 9 different scenarios.

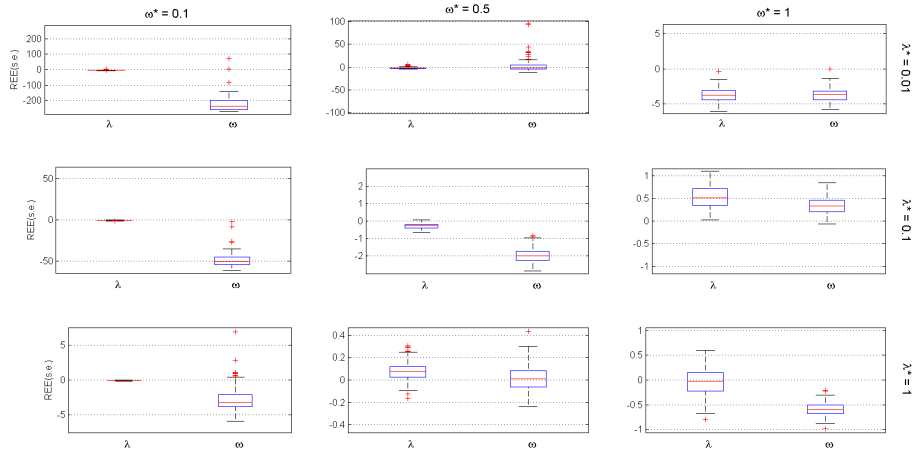

Figure 2: Relative estimation errors (in %) for the standard errors of $\hat{\lambda}$ and $\hat{\omega}$ obtained with 9 different scenarios.

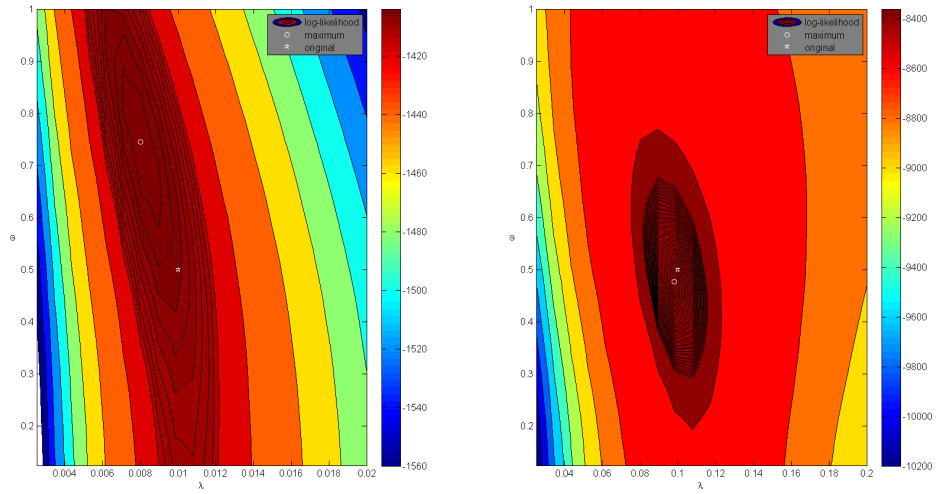

Figure 3: Observed log-likelihood as a function of λ and ω obtained with 2 different scenarios. Left: $(\lambda^\star, \omega^\star) = (0.01, 0.5)$, right: $(\lambda^\star, \omega^\star) = (0.1, 0.5)$.

Parameter	Estimates	Standard error	$P(\lvert\beta\rvert > \lvert\beta^{\text{obs}}\rvert)$
c_0	0.846	0.042	
$\beta_{0,d}$	-0.092	0.056	0.10
$\beta_{0,z}$	1.26	0.067	$<10^{-4}$
c_1	0.068	0.021	
$\beta_{1,d}$	0.002	0.027	0.94
$\beta_{1,z}$	0.009	0.040	0.81
c_2	0.0054	0.0022	
$\beta_{2,d}$	-0.0009	0.003	0.77
$\beta_{2,z}$	0.069	0.007	$<10^{-4}$
λ_0	0.0039	0.0011	
$\beta_{\lambda,d}$	0.002	0.42	0.999
α	1.64	0.11	
$\beta_{\alpha,d}$	-0.004	0.15	0.999
ω_0	0.431	0.022	
ω_1	0.196	0.011	
ω_2	0.0135	0.0015	
a	0.209	0.004	

Table 1: Primary biliary cirrhosis data: estimation of the population parameters.

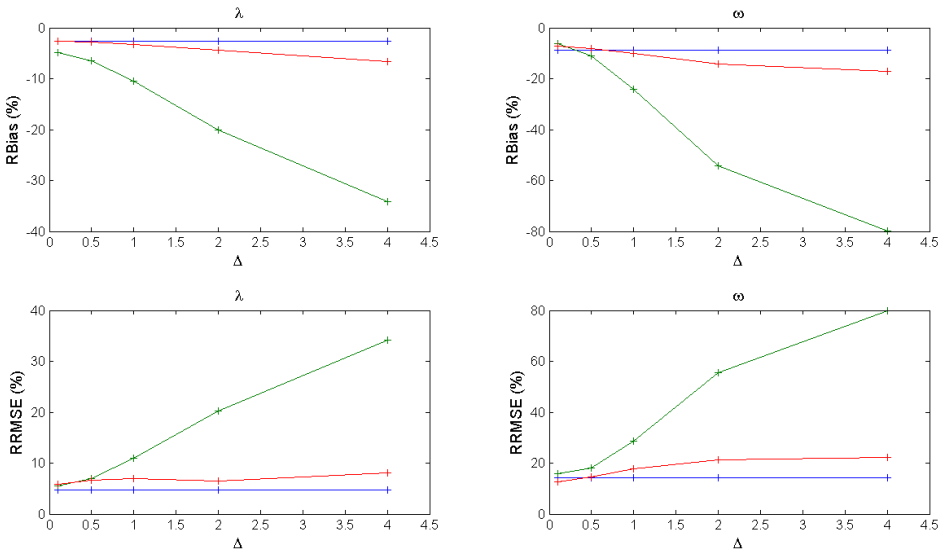

Figure 4: Relative bias and relative root mean square errors for λ and ω as a function of the width Δ of the censoring interval. Blue: using the exact event times; Red: taking correctly into account that the event is interval or right-censored; Green: taking incorrectly into account that the event is interval or right-censored.

Parameter	Estimates	Standard error	$P(\|\beta\| > \|\beta^{\text{obs}}\|)$
a	0.491	0.019	
b	0.463	0.022	
β_b	-0.239	0.054	<0.0001
c	0.097	0.013	
β_c	0.605	0.230	0.0076
ω_a	1.05	0.027	
ω_b	1.1	0.029	
$\rho_{a,b}$	0.889	0.009	

Table 2: Epileptic daily seizures count: estimation of the population parameters.

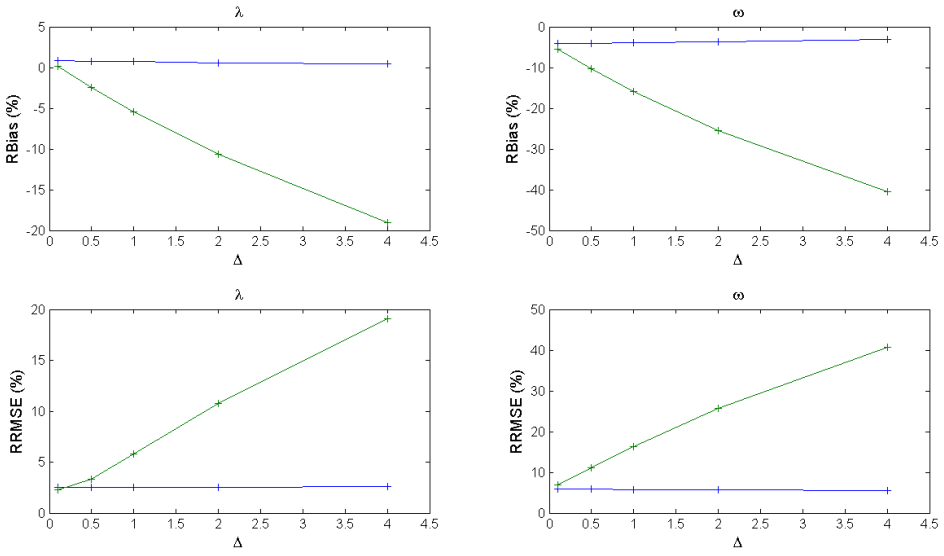

Figure 5: Relative bias and relative root mean square errors for λ and ω as a function of the width Δ of the censoring interval. Blue: taking into account the fact that the number of events is bounded ($k_{\max} = 5$); Green: ignoring the fact that the number of events is bounded.

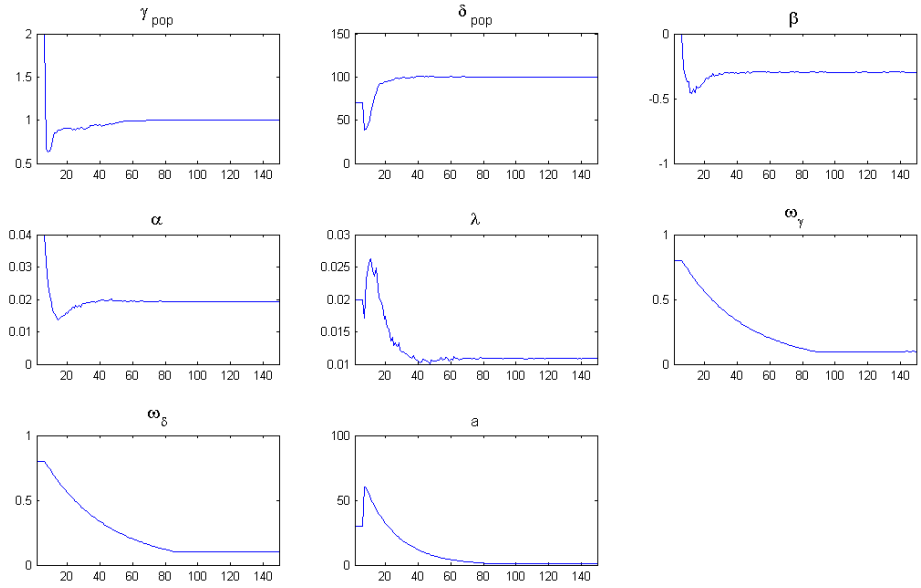

Figure 6: Convergence of the SAEM algorithm.

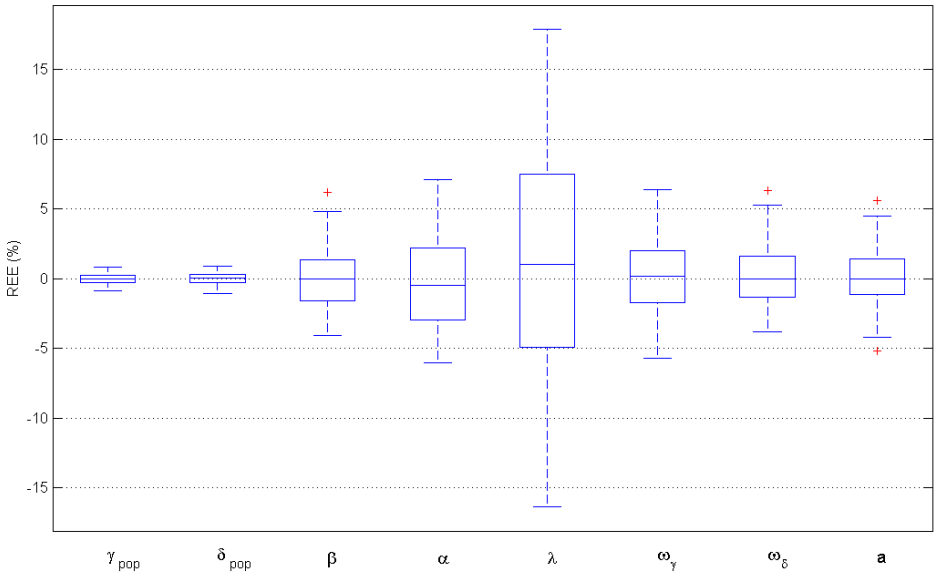

Figure 7: Relative estimation errors (in %) for the joint model.

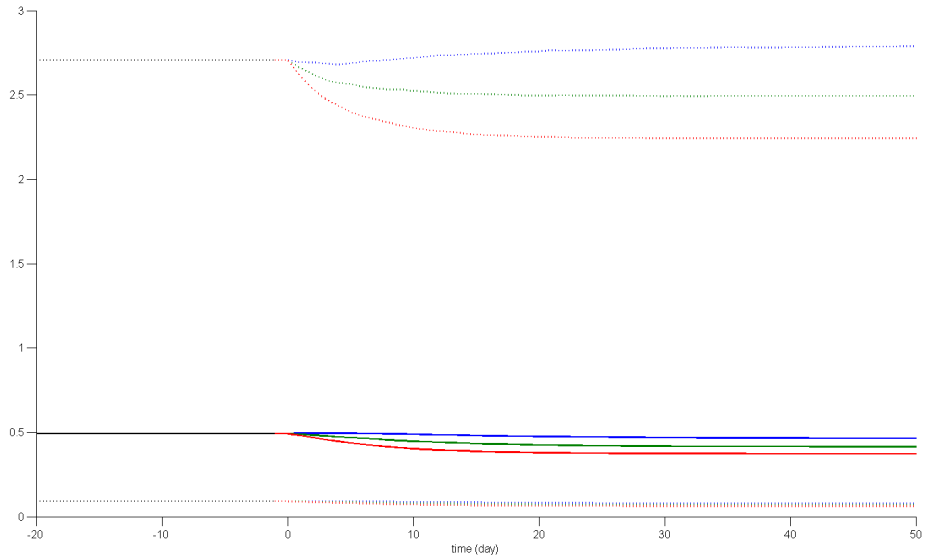

Figure 8: Hazard function for epileptic seizures count data. Hazard functions associated to different doses of gabapentin are displayed; blue: 0g (placebo); green: 0.6g; red: 1.5g. The median hazard functions are displayed with solid lines, the 90% prediction intervals with dotted lines.

Web-based Supplementary Materials "Joint modeling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the SAEM algorithm"

Cyprien Mbogning, Kevin Bleakley, and Marc Lavielle

December 23, 2013

1 Web Appendix A

General description of the SAEM algorithm.

Let $\theta^{(k-1)}$ denote the current estimate for the population parameters. Iteration k of the SAEM algorithm involves three steps [1, 2]:

- In the simulation step, $\theta^{(k-1)}$ is used to simulate the missing data $\psi_i^{(k)}$ under the conditional distribution $p(\psi_i|y_i, \theta^{(k-1)})$, $i = 1, \ldots, N$.
- In the stochastic approximation step, the simulated data $\psi^{(k)}$ and the observations y are used together to update the stochastic approximation $Q_k(\theta)$ of the conditional expectation $E\left(\log p(y, \psi; \theta)|y, \theta^{(k-1)}\right)$ according to:

$$Q_k(\theta) = Q_{k-1}(\theta) + \nu_k \left(\log p(y, \phi^{(k)}; \theta) - Q_{k-1}(\theta)\right), \qquad (1)$$

  where $(\nu_k)_{k>0}$ is a sequence of positive step sizes decreasing to 0 and starting with $\nu_1 = 1$.
- In the maximization step, an updated value of the estimate $\theta^{(k)}$ is obtained by maximization of $Q_k(\theta)$ with respect to θ:

$$\theta^{(k)} = \underset{\theta}{\operatorname{argmax}}\, Q_k(\theta).$$

This procedure is iterated until numerical convergence of the sequence $\left(\theta^{(k)}\right)_{k>0}$ to some estimate $\hat{\theta}$ is achieved. Convergence results can be found in [1].

When an estimate $\hat{\theta}$ has been obtained, estimates of the standard errors of its components can be derived by estimating the Fisher information matrix $I(\hat{\theta}) = -\partial^2 \log(p(y; \theta))/\partial\theta\partial\theta'|_{\theta=\hat{\theta}}$ following the stochastic approximation procedure suggested in [2], which requires simulation of the ψ_i's under $p(\cdot|y, \hat{\theta})$ via a Metropolis-Hastings algorithm.

Estimates of the ψ_i's can also be derived from the conditional distribution $p(\psi_i|y_i, \hat{\theta})$ such as the conditional mode or the conditional mean. Whatever the estimate chosen, simulating this conditional distribution via Metropolis-Hastings, or maximizing it, requires computing the conditional distribution of the observations $p(y_i|\psi_i)$.

Standard model selection criteria such as BIC require calculation of the observed log-likelihood $\log(p(y;\hat{\theta}))$. As the log-likelihood cannot be computed in a closed form here, it is approximated using an importance sampling procedure. This consists of drawing $\psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(M)}$ under a given sampling distribution $\tilde{\pi}$, and approximating the likelihood with:

$$p(y;\theta) \approx \frac{1}{M} \sum_{k=1}^{M} p(y|\psi^{(k)}) \frac{\pi(\psi^{(k)},\theta)}{\tilde{\pi}(\psi^{(k)})}.$$

Here also, we see that computation of $p(y_i|\psi_i)$ is required.

References

[1] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence Of A Stochastic Approximation Version Of The EM Algorithm. *The Annals Of Statistics*, 27:94–128, 1999.

[2] Estelle Kuhn and Marc Lavielle. Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.*, 49:1020–1038, 2005.

2 Web Appendix B

Computing the probability distribution for repeated time-to-events.

Exactly observed events

i) the last event is observed. Assume that we observe n events at times t_1, t_2, ..., t_n and that no event occurs after t_{end}. The vector of observations is $y = (t_1, t_2, \ldots, t_n)$ and

$$\begin{aligned} p(y|\psi) &= p(t_1, t_2, \ldots, t_n) \\ &= p(t_1|t_0)p(t_2|t_1)p(t_3|t_2)\ldots p(t_n|t_{n-1}). \end{aligned}$$

By definition, $p(t_j|t_{j-1}) = \lambda(t_j)e^{-\Lambda(t_{j-1},t_j)}$. Thus,

$$p(y|\psi) = \prod_{j=1}^{n} p(t_j|t_{j-1}) \tag{2}$$

$$= \prod_{j=1}^{n} \lambda(t_j)e^{-\Lambda(t_{j-1},t_j)}\,. \tag{3}$$

ii) the last event is not observed. Assume that we observe n events at times t_1, t_2, ..., t_n and that an event is known to occur at time $T_{n+1} > t_{end}$. Here, the vector of observations is $y = (t_1, t_2, \ldots, t_n, t_{n+1} > t_{end})$ and

$$\begin{aligned} p(y|\psi) &= p(t_1, t_2, \ldots, t_n)\mathbb{P}(T_{n+1} > t_{end}|T_n = t_n) \\ &= p(t_1|t_0)p(t_2|t_1)p(t_3|t_2)\ldots p(t_n|t_{n-1})\mathbb{P}(T_{n+1} > t_{end}|T_n = t_n) \\ &= \left(\prod_{j=1}^{n} \lambda(t_j)e^{-\Lambda(t_{j-1},t_j)}\right) e^{-\Lambda(t_n,t_{end})}\,. \end{aligned}$$

Single interval-censored events

Assume that n events occur between t_0 and t_{end} but that we only know that $t_1 \in [a_1, \ell_1]$, $t_2 \in [a_2, \ell_2]$, ..., $t_n \in [a_n, \ell_n]$.

i) the last event is observed. Here, no event occurs after t_{end}. The vector of observations is $y = (t_1 \in [a_1, \ell_1], t_2 \in [a_2, \ell_2], \ldots, t_n \in [a_n, \ell_n])$ and its joint probability distribution is:

$$\begin{aligned} p(y|\psi) &= \mathbb{P}(T_1 \in [a_1, \ell_1], T_2 \in [a_2, \ell_2], \ldots, T_n \in [a_n, \ell_n]) \\ &= \int_{a_1}^{\ell_1} \int_{a_2}^{\ell_2} \ldots \int_{a_n}^{\ell_n} p(t_1, t_2, \ldots, t_n)\, dt_1\, dt_2, \ldots, dt_n. \end{aligned}$$

Using equations 2-3,

$$\begin{aligned} p(t_1, t_2, \ldots, t_n) &= \prod_{j=1}^{n} \lambda(t_j) e^{-\Lambda(t_{j-1}, t_j)} \\ &= \left(\prod_{j=1}^{n} \lambda(t_j) \right) e^{-\sum_{j=1}^{n} \Lambda(t_{j-1}, t_j)} \\ &= \left(\prod_{j=1}^{n-1} \lambda(t_j) \right) \lambda(t_n) e^{-\Lambda(t_0, t_n)} \\ &= \left(\prod_{j=1}^{n-1} \lambda(t_j) \right) p(t_n|t_0). \end{aligned}$$

Thus, the multiple integral can be computed:

$$\begin{aligned} p(y|\psi) &= \left(\prod_{j=1}^{n-1} \Lambda(a_j, \ell_j) \right) \mathbb{P}(T_n \in [a_n, \ell_n] | T_{n-1} = t_0) \\ &= \left(\prod_{j=1}^{n-1} \Lambda(a_j, \ell_j) \right) \left(e^{-\Lambda(t_0, a_n)} - e^{-\Lambda(t_0, \ell_n)} \right). \end{aligned} \tag{4}$$

ii) the last event is not observed. Here, at least one event is known to occur after t_{end}. Thus, $y = (t_1 \in [a_1, \ell_1], t_2 \in [a_2, \ell_2], \ldots, t_n \in [a_n, \ell_n], t_{n+1} > t_{end})$. The previous result (see equation 4) holds with $a_{n+1} = t_{end}$ and $\ell_{n+1} = +\infty$:

$$\begin{aligned} p(y|\psi) &= \mathbb{P}(T_1 \in [a_1, \ell_1], T_2 \in [a_2, \ell_2], \ldots, T_n \in [a_n, \ell_n], T_{n+1} > t_{end}) \\ &= \left(\prod_{j=1}^{n} \Lambda(a_j, \ell_j) \right) e^{-\Lambda(t_0, t_{end})}. \end{aligned} \tag{5}$$

Multiple events per interval

Consider first a single interval $[0, \ell]$ and let $k_{\max}$ ($k_{\max} \leq +\infty$) be the maximum number of events. Let K be the number of events in $[0, \ell]$. For any $k < k_{\max}$, $K = k$ implies that the $(k+1)$-th event occurs after time ℓ. Then, for any $k < k_{\max}$,

$$\begin{aligned}
\mathbb{P}(K = k) &= \mathbb{P}(T_1 \in [0, \ell], \ldots, T_k \in [0, \ell], T_{k+1} > \ell;\ T_1 < \ldots < T_k < T_{k+1}) \\
&= \int_0^\ell \int_{t_1}^\ell \cdots \int_{t_{k-1}}^\ell \int_\ell^{+\infty} p(t_1, t_2, \ldots, t_k, t_{k+1})\, dt_1\, dt_2 \ldots dt_k\, dt_{k+1} \\
&= \int_0^\ell \int_{t_1}^\ell \cdots \int_{t_{k-1}}^\ell \int_\ell^{+\infty} \left(\prod_{j=1}^{k} \lambda(t_j) \right) p(t_{k+1}|t_0)\, dt_1\, dt_2 \ldots dt_k\, dt_{k+1} \\
&= \frac{\Lambda(0, \ell)^k}{k!} e^{-\Lambda(0,\ell)}. \qquad (6)
\end{aligned}$$

Remark. In the case of a constant hazard function $\lambda(t) = \lambda$, the inter-event times follow the exponential distribution with parameter λ. Then, the number of events in any interval of length ℓ follows a Poisson distribution with parameter $\Lambda(0, \ell) = \lambda\ell$. For any $k < k_{\max}$,

$$\mathbb{P}(K = k) = \frac{(\lambda\ell)^k}{k!} e^{-\lambda\ell}. \qquad (7)$$

Equation 6 thus shows that this type of property still holds for non-constant hazard functions $\lambda(t)$.

So, for a bounded number of events ($k_{\max} < +\infty$),

$$\begin{aligned}
\mathbb{P}(K = k_{\max}) &= 1 - \sum_{k=0}^{k_{\max}-1} \mathbb{P}(K = k) \\
&= 1 - \sum_{k=0}^{k_{\max}-1} \frac{\Lambda(0, \ell)^k}{k!} e^{-\Lambda(0,\ell)}. \qquad (8)
\end{aligned}$$

Consider now n contiguous intervals ($[\ell_{j-1}, \ell_j]; 1 \leq j \leq n$), where $\ell_0 = t_0$ and $\ell_n = t_{end}$. Let K_j be the number of events in interval $[\ell_{j-1}, \ell_j]$.

i) the last event is observed.. Let $s_{n-1} = \sum_{j=1}^{n-1} k_j$. Using equations 6 and 8, we can show that

$$\begin{aligned} p(y|\psi) &= \mathbb{P}(K_1 = k_1, K_2 = k_2, \ldots, K_n = k_{\max} - s_{n-1}) \\ &= \left(\prod_{j=1}^{n-1} \mathbb{P}(K_j = k_j) \right) \left(1 - \sum_{k=0}^{k_{\max}-s_{n-1}} \mathbb{P}(K_n = k) \right) \\ &= \left(\prod_{j=1}^{n-1} \frac{\Lambda(\ell_{j-1}, \ell_j)^{k_j}}{k_j!} e^{-\Lambda(\ell_{j-1},\ell_j)} \right) \\ &\qquad \times \left(1 - \sum_{k=0}^{k_{\max}-s_{n-1}} \frac{\Lambda(\ell_{n-1}, \ell_n)^{k}}{k!} e^{-\Lambda(\ell_{n-1},\ell_n)} \right). \end{aligned}$$

ii) the last event is not observed. This implies that the first non-observed event occurs after t_{end}. Using the above equation, it is straightforward to show that if $\sum_{j=1}^{n} k_j < k_{\max}$, then

$$\begin{aligned} p(y|\psi) &= \mathbb{P}(K_1 = k_1, K_2 = k_2, \ldots, K_n = k_n) \\ &= \prod_{j=1}^{n} \left(\frac{\Lambda(\ell_{j-1}, \ell_j)^{k_j}}{k_j!} e^{-\Lambda(\ell_{j-1},\ell_j)} \right). \end{aligned} \tag{9}$$

3 Web Appendix C

Joint modeling of PK and time-to-event data

The model. An anticoagulant is administrated to $N = 100$ patients during 7 days by combining oral and intravenous administrations. Adverse effects of this anticoagulant include hemorrhaging. We then measure for each patient the plasmatic concentration of the drug and the time of hemorrhages.

For the PK component of the model, we use a one compartment model and assume nonlinear elimination as given by the Michaelis-Menten equations:

$$\begin{aligned} \dot{A}_d(t) &= -k_a A_d(t) \\ \dot{A}_c(t) &= k_a A_d(t) - \frac{V_m\, A_c(t)}{K_m\, V + A_c(t)} \\ C(t) &= \frac{A_c(t)}{V}, \end{aligned}$$

where A_d (resp. A_c) is the amount in the depot (resp. central) compartment and C the concentration in the central compartment. The target compartments for the doses are A_d for oral administration and A_c for intravenous administration. The vector of PK parameters of the model is $\phi = (k_a, V, V_m, K_m)$.

The residual error model for the observed concentrations w_{ij} is a combination of a constant and a proportional error model:

$$w_{ij} = C(t_{ij}, \phi_i) + (a + b\, C(t_{ij}, \phi_i))\tilde{\varepsilon}_{ij}\ , \qquad \tilde{\varepsilon}_{ij} \sim_{i.i.d.} \mathcal{N}(0, 1).$$

We furthermore introduce a limit of quantification (LOQ) of 0.04 mg/l (this value was chosen in order to have about 10% of the concentrations left-censored, i.e. below the LOQ).

Considering hemorrhages as repeated events, we propose a model for the time to these events which assumes that the hazard depends on the concentration (risk of bleeding increases with concentration). We also assume that patients develop drug tolerance with time. We therefore propose a hazard model which combines a decreasing Weibull baseline and an increasing function of the predicted concentration:

$$\lambda_i(t) = \frac{\beta_i}{\gamma_i} \left(\frac{t}{\gamma_i}\right)^{\beta_i - 1} e^{\alpha_i\, C(t, \phi_i)}.$$

Hemorrhages are reported until $t = 200$ hours. Further events are treated as right-censored events.

The individual parameters $(ka_i, V_i, Vm_i, Km_i, \alpha_i, \beta_i)$ are assumed to be log-normally distributed and mutually independent. Parameter γ_i and the residual error parameters a and b are constant.

parameter ψ	population parameter ψ_{pop}	standard deviation ω_ψ
k_a	0.5	0.3
V	70	0.2
V_m	6	0.1
K_m	0.2	0.2
α	1	0.1
β	0.5	0.1
γ	15	0
a	0.05	0
b	0.05	0

A log-normal distribution for a parameter ψ means here that for any $i = 1, 2, \ldots N$,

$$\log(\psi_i) \sim_{i.i.d.} \mathcal{N}\left(\log(\psi_{\text{pop}}), \omega_\psi^2\right).$$

The design. We consider four arms which receive four different treatments, each treatment combining oral and intravenous (iv) delivery:

arm	size	amount (mg) oral	amount (mg) iv
1	25	100	50
2	25	100	25
3	25	50	50
4	25	50	25

The $N = 100$ patients receive an oral dose at the start of each of the first seven 7 days ($t = 0, 24, \ldots, 144$) and an iv dose at the midpoint of the day for 6 days, starting on day 2 ($t = 36, 70, \ldots, 156$). Concentrations are measured on day 1 at times 0.5, 4, 8, 12, 16, 20, 24, then only at the midpoint of the day for the next 5 days ($t = 36, 70, \ldots, 132$); then every four hours during day 7 ($t = 144, 148, 152, 156, 160$) and at times $t = 164, 168, 172$ during day 8.

Implemation of the model using MLXtran. Below is the MLXtran file used to encode the model for the observations, i.e. the structural model and the residual error model for the PK data and the hazard function for the time-to-events data. This file is used both for simulation and estimation using MONOLIX.

```
INPUT:
parameter =  {ka, V, Vm, Km, alpha, beta, gamma}

PK:
depot(type=1, target=Ad)
depot(type=2, target=Ac)

EQUATION:
ddt_Ad = -ka*Ad
ddt_Ac = ka*Ad - Vm/(V*Km+Ac)*Ac
C=Ac/V
if t<0
   lambda=0
else
   lambda=(beta/gamma)*(t/gamma)^(beta-1)*exp(alpha*C)
end

OBSERVATION:
Concentration = {type=continuous,prediction=C,error=combined1}
Hemorrhaging = {type=event,  hazard=lambda}
```

The predicted PK profile and hazard given by the model for one patient are displayed Figure 1, together with the simulated concentrations and events.
Estimation of the population parameters. Estimation of the population parameters is performed using SAEM in MONOLIX. Figure 2 displays the sequence of estimates $(\hat{\theta}_k)$ provided by SAEM.

Table 1 displays the true, initial and estimated values for each population parameter. Parameter estimation took 8 seconds, and Fisher matrix estimation 1 second.

Diagnostic plots. We briefly present in this section some basic diagnostic plots that can be used for model assessment. Obviously, it is expected here that we will obtain "very good" diagnostic plots since the data have been

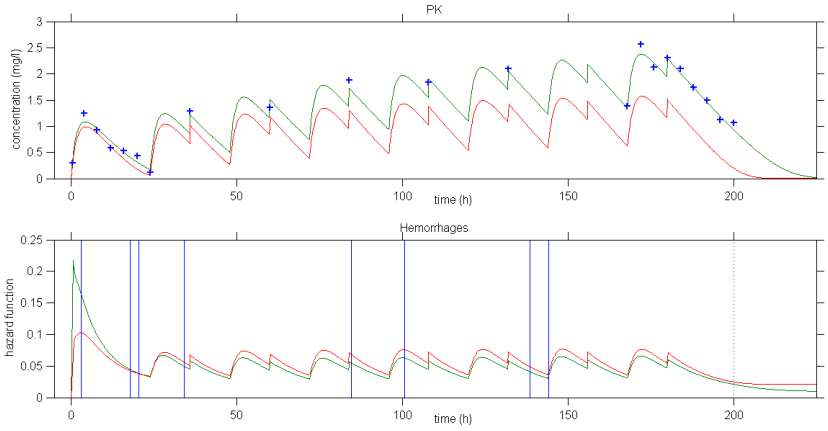

Figure 1: top: predicted and observed concentrations; bottom: hazard function and observed events. The individual model (computed with the individual parameters) is displayed in green, the population model (computed with the population parameters) is displayed in red, and observations are displayed in blue.

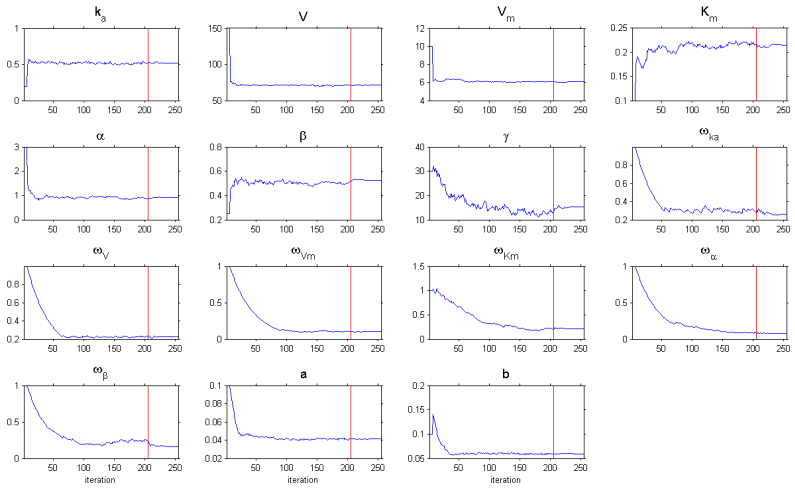

Figure 2: Sequence of estimates obtained with the SAEM algorithm. A constant stepsize $\nu_k = 1$ is used for 200 iterations (before the vertical red line), then ν_k decreases as $1/k$.

parameter θ	true value $\theta^\star$	initial value θ_0	estimated value $\hat{\theta}$	(s.e.)
ka_{pop}	0.5	0.25	0.517	(0.018)
V_{pop}	70	150	71.3	(1.7)
Vm_{pop}	6	10	6.06	(0.09)
Km_{pop}	0.2	0.1	0.215	(0.009)
α_{pop}	1	3	0.928	(0.075)
β_{pop}	0.5	0.25	0.523	(0.038)
γ_{pop}	15	30	15.2	(1.4)
ω_{ka}	0.3	1	0.259	(0.029)
ω_V	0.2	1	0.225	(0.017)
ω_{Vm}	0.1	1	0.104	(0.008)
ω_{Km}	0.2	1	0.208	(0.036)
ω_α	0.1	1	0.159	(0.044)
ω_β	0.1	1	0.077	(0.091)
a	0.05	0.1	0.042	(0.002)
b	0.05	0.1	0.059	(0.003)

Table 1: true values $\theta^\star$, initial values θ_0 used by SAEM, and estimated values $\hat{\theta}$ provided by SAEM.

simulated with the model we want to validate. All of these graphics are automatically produced by MONOLIX.

The graphs of observations vs. predictions (Figure 3) and of the individual weighted residuals (residuals obtained from the individual models, standardized, Figure 4) show that both the structural model and the residual error model properly fit the data.

Visual predictive checks (VPC) for the PK model and for the number of events are displayed Figures 5 and 6. The VPCs are stratified by arm (i.e. by treatment). These figures confirm that the joint model (including its statistical component) is able to reproduce the observed data.

A Monte-Carlo study. All these results so far have been obtained with *only one simulated dataset*. We now present a Monte-Carlo study to give stronger evidence that the SAEM algorithm is very efficient for estimating the population parameters in this complex joint model. We simulated 100 replicates of the same trial, using the same design and the same population parameters, and estimated the populations parameters for each simulated

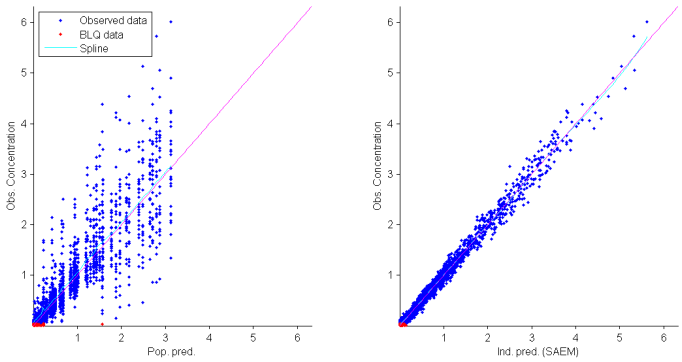

Figure 3: Observations vs predictions. Left: predictions given by the population model; right: predictions given by the individual models. Data below the limit of quantification are displayed in red.

trial. Initial values were randomly chosen around the true values. Figure 7 displays the distribution of the 15 estimated population parameters obtained with SAEM. We see that the PK parameters and the residual errors parameters can be estimated very accurately with this design. Estimation of the hazard model is less precise, mainly because the variability of the parameters α and β is difficult to estimate. Of course, it is not possible here to guess if the estimation errors are related to a purely statistical issue (statistical properties of the MLE are limited with a limited amount of information) or an issue with the algorithm (i.e. if SAEM does not converge properly to the MLE). Nevertheless, it is quite comforting to see that there is almost no bias and that most estimation errors are relatively small: we can reasonably conclude that SAEM "works quite well" in this complex situation.

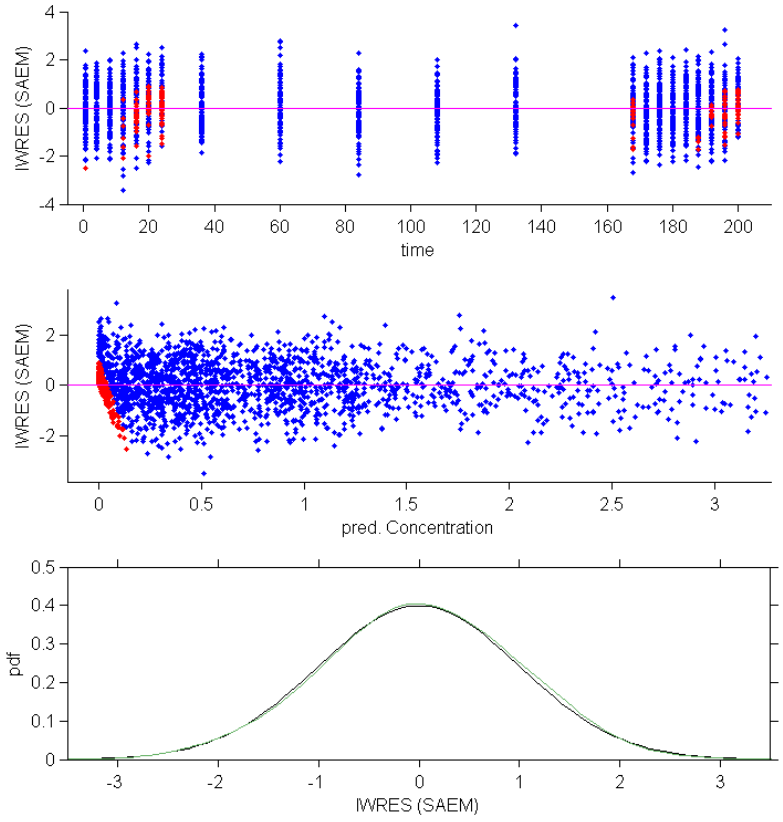

Figure 4: Individual weighted residuals (residuals obtained from the individual models and standardized). Top: residuals vs time; middle: residuals vs predictions; bottom: pdf of the estimated residuals (green) and pdf of the standardized normal distribution (black).

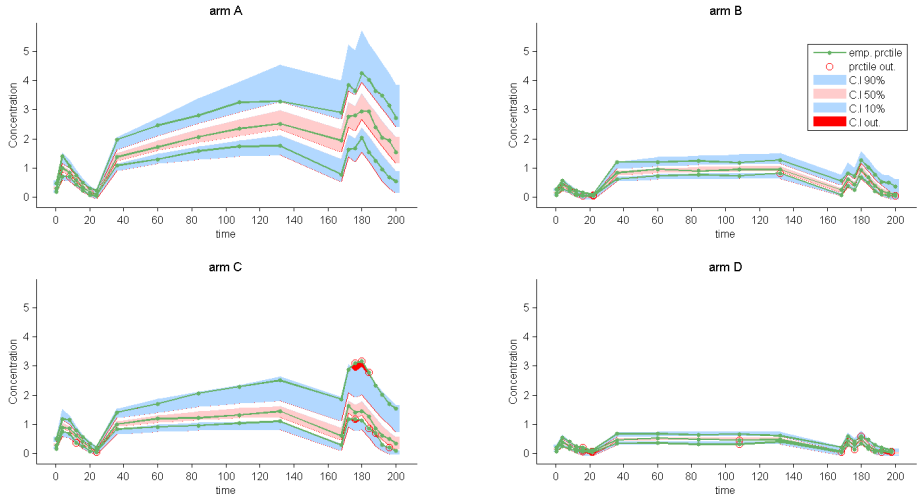

Figure 5: VPC for the PK model: for each arm of the trial, the empirical quantiles of order 10%, 50% and 90% of the PK data are compared with the 90% prediction intervals derived from the model for these quantiles.

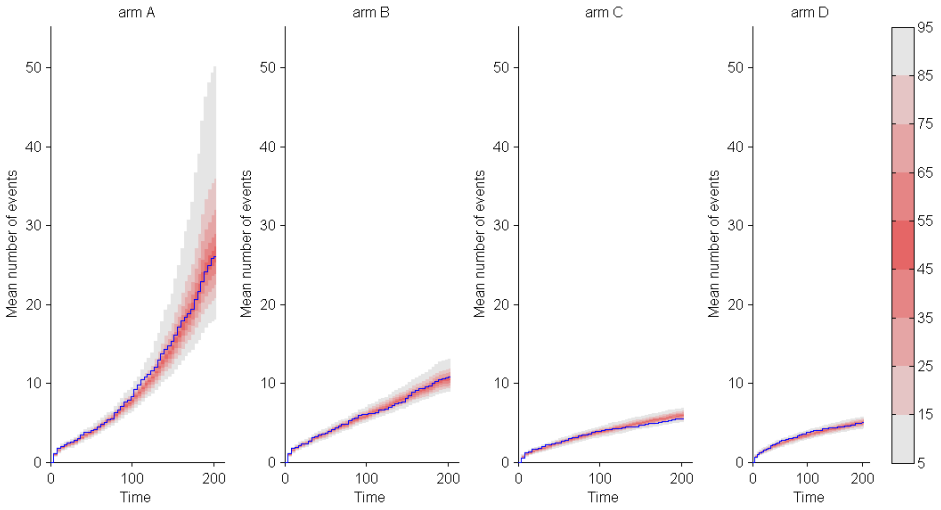

Figure 6: VPC for the number of events: for each arm of the trial, the empirical number of events is compared with the predicted distribution derived from the model for these quantiles.

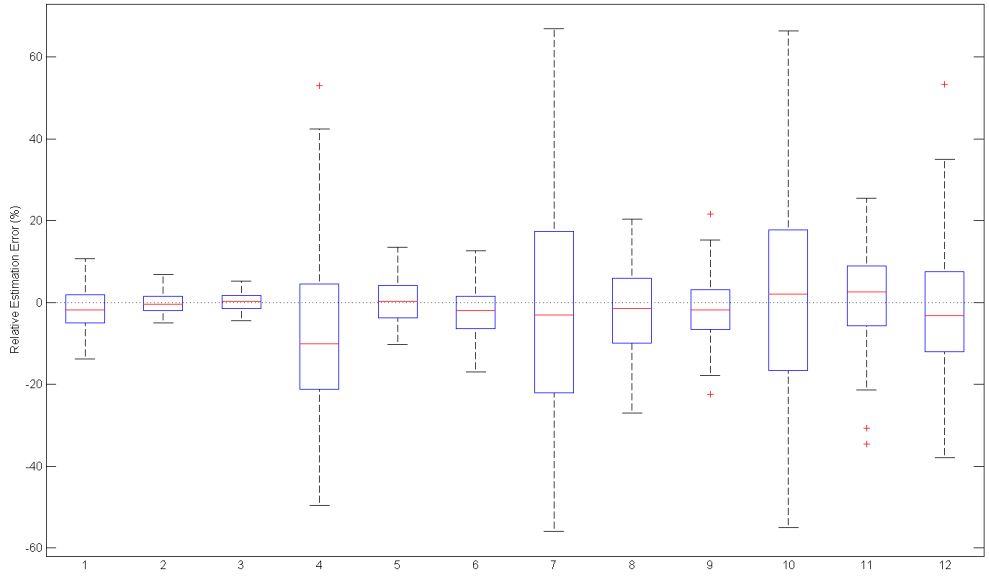

Figure 7: Empirical distribution of the relative estimation errors (in %) obtained from 100 simulated replicates of the trial. 1: ka_{pop}, 2: V_{pop}, 3: Vm_{pop}, 4: Km_{pop}, 5: α_{pop}, 6: β_{pop}, 7: γ_{pop}, 8: ω_{ka}, 9: ω_V, 10: ω_{Vm}, 11: ω_{Km}, 12: ω_α, 13: ω_β, 14: a, 15: b.