Towards using a full spectrum of early clinical trial data: a retrospective analysis to compare potential longitudinal categorical models for molecular targeted therapies in oncology

May 17th 2015

P. Colin[ab*], S. Micallef[a], M. Delattre[bc], P. Mancini[a] and E. Parent[bc]
[a]Biostatistique Oncologie, Sanofi R&D, Vitry-sur-Seine, France.
[b]AgroParisTech, UMR 518 MIA, 75005 Paris, France.
[c]INRA, UMR 518 MIA, 75005 Paris, France
*Pierre.Colin@sanofi.com - Biostatistique Oncologie, Sanofi R&D, 13 quai Jules Guesde, BP 14, 94403 Vitry-sur-Seine Cedex, France.

Abstract

Following the pattern of Phase I clinical trials for cytotoxic drugs, dose-finding clinical trials in oncology of molecularly targeted agents (MTA) aim at determining the Maximal Tolerated Dose (MTD). In classical phase I clinical trials, MTD is generally defined by the number of patients with short-term major treatment toxicities (usually called dose-limiting toxicities, DLT), occurring during the first cycle of study treatment (e.g. within the first 3 weeks of treatment). However, S. Postel-Vinay (2011) highlighted that half of grade 3 to 4 toxicities, usually considered as Dose Limiting Toxicities, occur after the first cycle of MTA treatment. In addition, MTAs could induce other moderate (e.g. grade 2) toxicities which could be taken into account depending on their clinical importance, chronic nature and duration. Ignoring these late toxicities may lead to an underestimation of the drug toxicity and to wrong dose recommendations for phase II and III clinical trials. Some methods have been proposed, such as the Tite-CRM (Cheung 2000 and Mauguen 2011), to take into account the late toxicities. We suggest approaches based on longitudinal models (Doussau 2013). We compare several models for longitudinal data, such as transitional or marginal models, to take into account all relevant toxicities occurring during the entire length of the patient treatment (and not just the events within a predefined short-term time-window). These models allow the statistician to benefit from a larger amount of safety data which could potentially improve that accuracy in MTD assessment.

Keywords: Late Toxicities; Longitudinal data; Dose-Finding; Oncology

1 Introduction & Motivation

The final goal of early phase oncology trials consists in choosing a Recommended Phase 2 Dose (RP2D) among a set of different explored doses. Selecting the RP2D is of major interest since it conditions the following clinical development phases[1, 2, 3]. On top of the importance of the task, recommending a dose is difficult for many reasons: limited amount of data, heterogeneous population, complex mechanism of distribution and target reaching. Because of study duration constraints or incompleteness of the data, designs generally rely on early acute toxicity but final dose recommendation should handle any kind of clinical meaningful toxic event and not just the early acute adverse events.

In this paper, we propose an approach reassessing the Maximum Tolerated Dose (MTD) at the end of the trial when most of the late toxicity data are observed. We think that the final recommended dose should not necessarily be the latest explored one, as it is the case with classical design. Indeed, the latest explored dose usually corresponds to the one closest of a targeted rate of early severe toxicity [5], or to the highest limiting early acute toxicity risks [6, 7] but is not necessarily admissible in terms of overall safety.

For more than a decade, targeted therapies have been appearing with complex mechanisms of action targeting specific pathways, and with complex pharmacokinetic profiles. These Molecular Targeted Agents (MTA) do not aim at eradicating tumor cells in a short time frame as cytotoxic treatments do, but rather they try to continuously repair the pathological behaviour of tumor cells. Usually, they are administered in a chronic fashion with a high frequency (weekly or even daily). These specificities lead to a change in the nature of related toxic events and therefore, potentially, in the adverse events limiting the dose[1, 8]. In particular, severe late onset toxicities could appear due to drug accumulation following a chronic schedule. Prolonged or repeated moderate toxicity are also relevant for consideration: some moderate toxicities, although deemed to be tolerable when single, significantly impact patient health when becoming chronic, acquiring an intolerable character [8, 9]. Therefore, these late onset toxicities and moderate grade repeated adverse events should be considered when MTD is established. When dealing with compounds causing such toxicities, MTD assessment based on binary endpoint of toxicity occurring within a short time period leads to the recommendation of a final dose regimen which is not tolerated well in the long-term.

Therefore, for MTA, even if early Dose Limiting Toxicities (DLT) are still largely used for driving dose escalation, experts have come to an agreement that, at the end of the study, a dose should be recommended on the basis of any major toxic events whatever the time window within which they appear[9]. Models are then needed to predict late, repeated or prolonged events with their severity to help in the assessment of the recommended dose, and thus process all information conveyed by the data.

In this paper, we consider the position of data prediction for dose recommendation at the end of study when no more data are awaited. The purpose is not to try to deal with incompleteness of the data as in an ongoing study framework but rather to propose models able to predict any relevant safety data with their time of occurrence as a function of dose for final dose recommendation purposes.

Different approaches have already been proposed for working with late toxicity data, even if time is still needed to translate this research based knowledge into operational practice. Time-to-Event Continuous Reassessment Method (TITE-CRM) [10] is probably the simplest of the proposed methods. It was the first attempt to incorporate information coming from late event in partially observed subjects. This method consists in using a binomial likelihood weighted according to the proportion of observation duration. Another interesting method is the time to event approach. Commonly used to model efficacy data, the time to event approach has also been used for safety data in the context of dose-finding trials [11]. A 2-parameter Weibull distribution is used to model the occurrence of DLT at any time during study with a DLT log-hazard function increasing linearly with the dose. Toxicity scores (TS) is another approach which consists in defining and modelling a score describing the intensity of the cumulated toxicity events observed in patients. By setting a target score, MTD is assessed. This approach is derived in different ways by considering: 1. different toxicity score functions, such as the toxicity index proposed by Rogatko [12], the Equivalent Toxicity Level [13] or the toxicity burden score [14] 2. different regression models, such as the probit model [15], the Beta regression model [16], or the Multinomial regression [17]. TS methods allow an assessment of the toxicity load taking into account the type and the severity of the adverse events regardless of the time they occur, since a toxicity score function can be elicited.

Even if all of these methods are interesting for MTD assessment, only the model proposed by Doussau et. al [18] is able to model at the same time, the severity of an adverse event, the time to the occurrence of the event and its duration. TITE-CRM and TS methods do not allow time to event occurrence to be modelled. TITE-CRM approach and time to event methods do not allow the severity of the events to be modelled.

In this work, we propose to explore different marginal and Markov models able to predict longitudinal categorical data as is typically observed in early phase trials for the final assessment of MTD. Both population-averaged and mixed-effect versions of these models are investigated. In the next section, the statistical models under assessment will be described. Section 3 reports the simulation plan allowing the comparison of the competing models and the corresponding results. Finally, we discuss in Section 4 how the retained models can help in the recommendation of a dose at the end of a typical phase I trial.

2 Statistical models

Let N be the number of patients included in the clinical trial, and let Y_{ij} be the random variable of toxicity observation for individual $i = 1, \ldots, N$ at j^{th} time point after drug intake, $j = 1, \ldots, n_i$. For the sake of simplicity, the Y_{ij}'s are assumed regularly spaced in time. Let $Y_i = (Y_{i1}, \ldots, Y_{i,n_i})'$, where n_i is the total number of observations for patient i. Note that the number of individual toxicity observations n_i is not necessarily the same for all patients, typically when early withdrawals occur during the clinical trial. Most often when dealing with toxicities, Y_{ij} is an ordered categorical variable taking integer values from 0 (no severe toxicity events) to K (most severe toxicity events) corresponding to a grading of the event severity. Implicitly, an initial observation at time 0 (just before drug intake) Y_{i0} is assumed, but not measured. It will be considered hereinafter that $Y_{i0} \equiv 0$ for all $i = 1, \ldots, N$. This implies that immediate toxic reactions at very first drug administration will only be explained by baseline covariates, since treatment duration and cumulative drug exposure will be both close to zero. Hereafter, such an assumption on Y_{i0} will be especially useful for Markov approaches for full specification of the corresponding models. Let $x_{ij} = (x_{ij}^1, \ldots, x_{ij}^p) \in \mathbb{R}^p$ denote some possibly time-varying explanatory variables for individual i (such as the time to the beginning of the treatment of the j^{th} record) and let $x_i = (x_{ij})_{j \in 1,\ldots,n_i}$ be the matrix of explanatory variables of subject i for the whole observation period with possibly subject-specific parameter φ_i. The N patients are reasonably assumed to be independent, thus two observation variables $Y_{ij}\,|x_i, \varphi_i$ and $Y_{i'j'}\,|x_{i'}, \varphi_{i'}$ would be considered independent when $i \neq i'$ hereinafter. The objective of the present section is to propose statistical models for the Y_{ij}'s which adequately describe and explain the variability in the occurrence of toxic effects. As Figure 1 exemplifies, this variability is typically split into some between-patient variability (*ie* some patients are more susceptible to the toxic effects than others) and some within-patient variability (*ie* toxicity manifestations for a given patient vary over time). Specifically, toxic effects induced by MTA-based treatments turn out to be persistent, suggesting some dependence between the observations of a single patient. The graphs in Figure 1 show that when a toxic effect appears, it is more likely to persist several weeks. The usual statistical models proposed for the analysis of early clinical trials in oncology are not longitudinal data models, and therefore, they do not account for this dependence between the observations of each patient. Doussau *et al*[18] proposed a mixed-effect categorical model for toxicity data. Several modelling approaches are suggested in what follows through either population-averaged models or random effects models. For the sake of simplicity, these modelling approaches are presented in detail for the simplest case of binary responses ($K = 1$) but they could be easily extended to any value of K as briefly shown in subsection 2.2.

2.1 Models for binary longitudinal data

2.1.1 Population-averaged models

In population-averaged models, the relation between the observations and the subject-specific explanatory variables is assumed to be the same for all subjects. This relation is entirely parametrized by the set of parameters $\varphi \in \mathbb{R}^q$ which is common to the N subjects (this is equivalent to setting $\varphi_i \equiv \varphi$ for all individuals $i \in 1, \ldots, N$ in the population). With such an approach, the variability between subjects is supposed to be fully (and only) described by means of the explanatory variables. Two population averaged modelling strategies are mainly referenced in the literature to model longitudinal data: marginal models and transition models (see for instance [19] and [20]). The use of either marginal or transition models depends on the main scientific question in hand and corresponds to different handlings of the within-subject variability in the data, as we shall see hereafter. Thus, the major difference between these two modellings lies in the interpretation of the regression coefficients.

a) Markov models

Focusing on time transitions, Markov models are a possible way to describe longitudinal data: one is then interested in how some explanatory variables influence the change of toxicity level over time. The idea of parsimony behind these models is that what will happen in the future is only conditioned by the present state of the system, not by its entire past history. Since some toxic effects might be persistent in time, this seems to be a reasonable assumption in oncology. Here, one is interested in toxic effects induced by a given treatment. Since according to our notations $j = 0$ matches the beginning of the treatment period, it can reasonably be assumed that treatment-related toxic effects are never recorded at time t_{i0}, thus the following known initial state for the Markov chain:

$$\mathbb{P}(Y_{i0} = 0) = 1. \tag{1}$$

Extensions to any other (possibly φ-parametrized) initial distributions would naturally be straightforward. Let us denote $H_{ij} = (Y_{i,j-1}, \ldots, Y_{i1})$ the past recorded trajectory until record Y_{ij} with the convention that $H_{i1} = \emptyset$. Additive time homogeneous transition models could be specified in the following general form:

$$\mathbb{P}(Y_{ij}=1|H_{ij},\varphi)=F(\mu+\sum_{\ell=1}^{p}\beta_\ell x_{ij}^{\ell}+f(H_{ij},\alpha)), \tag{2}$$

where

$$\begin{aligned} f(H_{ij},\alpha) &= \sum_{r=1}^{\min(q,j)}\alpha_r \mathbf{1}(Y_{i,j-r}=0) \quad \text{if } j>1, \\ f(H_{i1},\alpha) &= 0. \end{aligned} \tag{3}$$

Here, parameters μ, β_ℓ and α_r are unknown constants making up the parameter set φ, F stands for any cumulative distribution function, and q is the order of the Markov chain. First-order Markov models ($q=1$) are the most widely used transition models. By noting $p_{ij}^{\varphi}(1|k')$ the probability $\mathbb{P}(Y_{ij}=1|Y_{i,j-1}=k',\varphi)$, $k'=0,1$, equation (3) becomes

$$\begin{aligned} p_{ij}^{\varphi}(1|0) &= F(\mu+\textstyle\sum_{\ell=1}^{p}\beta_\ell x_{ij}^{\ell}+\alpha_1) \quad , \\ p_{ij}^{\varphi}(1|1) &= F(\mu+\textstyle\sum_{\ell=1}^{p}\beta_\ell x_{ij}^{\ell}), \end{aligned} \tag{4}$$

the transition matrix from time $j-1$ to time j of subject i can be easily derived as:

$$Q_{ij}^{\varphi}=\begin{pmatrix} 1-p_{ij}^{\varphi}(1|0) & p_{ij}^{\varphi}(1|0) \\ 1-p_{ij}^{\varphi}(1|1) & p_{ij}^{\varphi}(1|1) \end{pmatrix} \tag{5}$$

and the full set of (unknown) model parameters is given by $\varphi=(\mu,\beta_1,\ldots,\beta_p,\alpha_1)$.

In transition models defined through equation (2), the joint distribution of the data vectors Y_i can be easily derived through the Markov property that lends a parsimonious form to the general decomposition:

$$\mathbb{P}(Y_i|\varphi)=\mathbb{P}(Y_{i0}|\varphi)\prod_{j=1}^{n_i}\mathbb{P}(Y_{ij}\,|H_{ij},\varphi). \tag{6}$$

Thus, likelihood-based methods as well as moment-based ones can be implemented to estimate the unknown parameter φ from the observations. In particular, the reader's attention is drawn to the fact that here, equation (2) specifies a Generalized Linear Model (GLM) for the conditional distribution of Y_{ij} given the past responses H_{ij}. Indeed, equation (2) regresses Y_{ij} on an extended set of explanatory variables which combines the x_{ij}'s and H_{ij}. Thus, it is possible to proceed with estimation of the transition parameters φ using inferential techniques specific to GLMs for independent data. Since the resolution of the estimating equations is generally intractable in GLMs, some numerical tricks are required. The well-known Iterative Weighted Least Squares (IWLS) method is such a technique [21]. It is implemented in the standard `glm` function of the `R` software.

b) Marginal models with incomplete specification

When choosing marginal models, by definition, one focuses on the potential effect of explanatory variables on the marginal distributions of the individual temporal observations Y_{ij} (theoretically obtained after integrating out all other $Y_{ij'}$, $j' \neq j$ in equation (6)). Marginal models are usually stated in two steps: the marginal distributions of the Y_{ij}'s are modelled separately from the within-subject correlations. In a very general setting, marginal models could be written as follows:

- *Step 1:* specification of the marginal distribution of each variable Y_{ij} as a function of explanatory variables

$$\mathbb{P}(Y_{i,j} = 1|\varphi) = F(\mu + \sum_{\ell=1}^{p} \beta_\ell x_{ij}^\ell) \ , \ 1 \leq i \leq N \ , \ 1 \leq j \leq n_i, \tag{7}$$

- *Step 2:* specification of the correlation structure between two toxicity observations Y_{ij} and Y_{ik} from the same subject

$$\mathbf{Corr}(Y_{i,j}, Y_{i,k}|\varphi) = \rho_{jk}(\alpha) \ , \ 1 \leq i \leq N \ , \ 1 \leq j, k \leq n_i. \tag{8}$$

Here, F could be any cumulative distribution function, parameters μ and β_ℓ are constants to be estimated, and ρ_{jk} is a known function of an unknown set of parameters α which specifies the nature of the autocorrelation. One specific example might be the autoregressive autocorrelation model:

$$\rho_{jk}(\alpha) = \alpha^{|j-k|} \ , \ \alpha \in [0, 1], \tag{9}$$

which assumes that measurements spaced in time are positively correlated, but less correlated than measurements close in time. Many other autocorrelation models can be chosen depending on the user's assumptions concerning the dependence between individual observations. Again, setting $\beta = (\beta_1, \ldots, \beta_p)$, the full set of (unknown) model parameters is given by $\varphi = (\mu, \beta, \alpha)$. Note that, although similar notations are used for the parameters as in the Markovian case, they do not entail the same interpretation (see subsection 2.3 below).

In addition, with the most remarkable exception of independence $\mathbb{P}(Y_i|\varphi) = \prod_{j=1}^{n_i} \mathbb{P}(Y_{ij}|\varphi)$, equations (15) and (16) are not sufficient to specify the entire model properly. Indeed, it is generally not possible to derive the joint distribution of the whole vector of individual observations Y_i from (15) and (16). Thus, likelihood methods cannot be implemented to estimate the unknown parameters φ from the observations. For this reason, a well-known alternative comes with the resolution of Generalized Estimating Equations (GEE), see for example [22] for details. Marginal models also fall within the scope of marginal quasi-likelihood and penalized quasi-likelihood methods introduced in [24] and [25] respectively.

2.1.2 Random effects models

In many cases however, the knowledge of the x_{ij}'s is not sufficient to fully describe the inter-patient variability. In such cases, it is preferable to adopt a subject-specific approach in which the responses (*ie* the temporal toxicity observations) are modelled as a function of covariates and subject-specific random effects φ_i. Such models are usually defined hierarchically by defining the φ_i's as random variables, so that population-averaged conclusions can still be derived through the so-called population parameters θ. Defining the individual parameters as random variables is also a way to handle the correlations between temporal observations from each subject. Random-effects models are specified in two stages:

- *Stage 1*: the conditional distribution of the individual observations Y_i given the individual parameters φ_i assumes the same kind of relation between the temporal observations and the explanatory variables for the N subjects:
$$Y_i|\varphi_i, x_i \sim p(\cdot|\varphi_i, x_i), \tag{10}$$
- *Stage 2*: the individual parameters φ_i are defined as independent random variables, the distribution of which is parametrized by a set of population parameters θ:
$$\varphi_i \underset{i.i.d.}{\sim} p(\cdot|\theta). \tag{11}$$

In most cases, the individual parameters are assumed to be Gaussian random variables. Multiple variations of the random-effects model could be derived for longitudinal categorical toxicity observations. Two specific examples of interest among so many others are given here:

1. The *generalized linear mixed-effects model*:
$$\begin{aligned}\mathbb{P}(Y_{i,j} = 1|\varphi_i) &= F(\mu_i + \sum_{\ell=1}^{p} \beta_{i,\ell} x_{ij}^{\ell}) \ , \\ \mathbb{P}(Y_{i,j}, Y_{i,j'}|\varphi_i) &= \mathbb{P}(Y_{i,j}|\varphi_i) \times \mathbb{P}(Y_{i,j'}|\varphi_i) \ , \\ \varphi_i = (\mu_i, \beta_{i,1}, \dots, \beta_{i,p})' &\underset{i.i.d.}{\sim} \mathcal{N}(\varphi, \Omega).\end{aligned} \tag{12}$$
2. The *mixed-effects first-order Markov model*, an extension of the above transition models (2) - (3) in a random-effects approach:
$$\begin{aligned}\mathbb{P}(Y_{ij} = 1|Y_{i,j-1}, \varphi_i) &= F(\mu_i + \sum_{\ell=1}^{p} \beta_{i,\ell} x_{ij}^{\ell} + \alpha_i \mathbf{1}(Y_{i,j-1} = 0)), \\ \varphi_i = (\mu_i, \alpha_i, \beta_{i1}, \dots, \beta_{ip})' &\underset{i.i.d.}{\sim} \mathcal{N}(\varphi, \Omega).\end{aligned} \tag{13}$$

In both examples, the population parameters are taken as the mean and the covariance matrix of normal population distribution $\theta = (\varphi, \Omega)$.

Inference of the population parameters θ in mixed-effects models is classically performed by maximization of the model likelihood. Owing to their hierarchical definition through equations (10) and (11), the joint marginal likelihood of each individual sequence of observations is straightforwardly given by integrating out $p(\cdot, x_i|\varphi_i)$ over the distribution of the random effects:

$$\mathbb{P}(Y_i|\theta) = \int p(Y_i|\varphi_i, x_i) p(\varphi_i|\theta) d\varphi_i. \tag{14}$$

In many cases, equation (14) does not have any explicit expression. Thus, maximum likelihood is often performed based on linearizations of (14) (see for instance chapter 6 in [26]), numerical approximations of the integral or via EM-based algorithms[23]. Marginal quasi-likelihood [24] and penalized pseudo-likelihood methods [25] have also been specifically developed to perform parameter estimation in mixed-effects generalized linear models. As in the fixed-effects setting, by definition of the Markov models given above, mixed-effects Markov models fall within the scope of this method. To fit generalized linear mixed models, one could use the function `glmer` of the `R` software, which approximates the integral in (14) by quadrature methods.

2.2 Models for ordinal longitudinal data

In practice, toxicity data may fall into more categories than just two possibilities, the possible toxicity values establishing a ranking between symptoms (see Figure 1 for instance). Some Markov and marginal models for longitudinal ordinal data could be easily derived from the longitudinal models for binary data presented in Section 2.1 whether in a fixed-effects setting or in a random effects setting. Without loss of generality, it is considered here that the grades of toxicity are integer values from 0 (no severe toxic event) to K (the most severe toxic events). Some possible models for this setting are listed below, but not detailed. Indeed, their interpretation and inference are basically similar to those for the corresponding binary data models.

(a) *Fixed-effects (population averaged) Markov models:*

$$\begin{array}{rll} \mathbb{P}(Y_{ij} \leq k|H_{ij}, \varphi) & = F(\mu_k + \sum_{\ell=1}^{p} \beta_\ell x_{ij}^\ell + f(H_{ij}, \alpha)) & 0 \leq k \leq K-1, \\ \mathbb{P}(Y_{ij} \leq K|H_{ij}, \varphi) & = 1, & \\ f(H_{ij}, \alpha) & = \sum\limits_{r=1}^{\min(q,j)} \sum\limits_{k=0}^{K-1} \alpha_{rk} \mathbf{1}(Y_{i,j-r} = k) & \text{if } j > 1, \\ f(H_{i1}, \alpha) & = 0 & \text{otherwise,} \end{array}$$

where parameters $\mu_0 < \mu_1 < \ldots < \mu_{K-1}$, β_ℓ and α_{rk} are unknown constants and compose the parameter set φ, F stands for any cumulative distribution function, and q is the order of the Markov chain. As above, for $q = 1$ (first-order Markov models), the transition matrix from time $j-1$ to time j of subject i can be easily derived from the probabilities $\mathbb{P}(Y_{ij} \leq k | H_{ij}, \varphi)$.

(b) *Fixed-effects (population-averaged) marginal models with incomplete specification:* As above, marginal models for longitudinal ordinal data are defined in two steps:

- *Step 1:* specification of the marginal distribution of each variable Y_{ij} as a function of explanatory variables

$$\begin{array}{rcll} \mathbb{P}(Y_{i,j} \leq k|\varphi) & = & F(\mu_k + \sum_{\ell=1}^{p} \beta_\ell x_{ij}^\ell) & , \ 0 \leq k \leq K-1, \\ \mathbb{P}(Y_{i,j} \leq K|\varphi) & = & 1 & , \ 1 \leq i \leq N \ , \ 1 \leq j \leq n_i, \end{array} \tag{15}$$

- *Step 2:* specification of the correlation structure between two toxicity observations Y_{ij} and $Y_{ij'}$ from the same subject

$$\mathbf{Corr}(Y_{i,j}, Y_{i,j'}|\varphi) = \rho_{jj'}(\alpha) \ , \ 1 \leq i \leq N \ , \ 1 \leq j, j' \leq n_i. \tag{16}$$

F stands for any cumulative distribution function, $\varphi = (\mu_0, \ldots, \mu_{K-1}, \beta_1, \ldots, \beta_p, \alpha)$ is the unknown vector of model parameters to be estimated, such that $\mu_0 < \mu_1 < \ldots < \mu_{K-1}$, and $\rho_{jj'}$ is a known function of an unknown set of parameters α which specifies the nature of the autocorrelation, for example an autoregressive one.

(c) *Generalized linear mixed-effects models*:

$$\begin{aligned} \mathbb{P}(Y_{i,j} \leq k|\varphi_i) &= F(\mu_{ik} + \sum_{\ell=1}^{p} \beta_{i,\ell} x_{ij}^\ell) \ , \ 0 \leq k \leq K-1, \\ \mathbb{P}(Y_{i,j} \leq K|\varphi_i) &= 1 \ , \\ \mathbb{P}(Y_{i,j}, Y_{i,j'}|\varphi_i) &= \mathbb{P}(Y_{i,j}|\varphi_i) \times \mathbb{P}(Y_{i,'j}|\varphi_i) \ , \\ \varphi_i = (\mu_{i0}, \ldots, \mu_{i,K-1}, \beta_{i1}, \ldots, \beta_{ip})' &\underset{i.i.d.}{\sim} \mathcal{N}(\varphi, \Omega), \end{aligned} \tag{17}$$

where F is a given cumulative distribution function and $\theta = (\varphi, \Omega)$ is the set of population parameters.

(d) *Mixed-effects Markov models:*

$$\begin{aligned}
\mathbb{P}(Y_{ij} \leq k | H_{ij}, \varphi_i) &= F(\mu_{ik} + \sum_{\ell=1}^{p} \beta_{i,\ell} x_{ij}^{\ell} + f(H_{ij}, \alpha_i)) \ , \ 0 \leq k \leq K-1, \\
\mathbb{P}(Y_{ij} \leq K | H_{ij}, \varphi_i) &= 1, \\
f(H_{ij}, \alpha_i) &= \sum_{r=1}^{j-1} \sum_{k=0}^{K-1} \alpha_{irk} \mathbf{1}(Y_{i,j-r} = k), \\
\varphi_i = (\mu_{i0}, \ldots, \mu_{i,K-1}, \beta_{i1}, \ldots, \beta_{ip}, \alpha_i)' &\underset{i.i.d.}{\sim} \mathcal{N}(\varphi, \Omega), \qquad (18)
\end{aligned}$$

where, as above, F is a given cumulative distribution function and $\theta = (\varphi, \Omega)$ is the set of population parameters.

2.3 Some examples

To cast light on the difference between the previous model families (Markov and marginal models whether in a population-averaged or a random-effects setting), as well as on the nature of this difference, a few model examples are presented and compared. For the sake of simplicity, the toxicity observations are considered as binary outcomes. This would correspond to cases where the user's interest would be for instance the presence ($Y_{ij} = 1$) or absence ($Y_{ij} = 0$) of toxic effects over time or correspond to the appearance with time of disabling ($Y_{ij} = 1$)/non-disabling ($Y_{ij} = 0$) toxic effects. Once again, in practice, there may be more categories than just two possibilities, but the binary situation will capture the essence of the example, whilst avoiding arithmetic and notational complexities. For the purpose of the example here, only one time non-varying explanatory variable c_i is considered. The main focus will be on the meaning of the regression coefficient β related to covariate c_i. For that purpose, the expression of the marginal distribution of the temporal observations with respect to the model parameters in each example - entirely given by the probability $\mathbb{P}(Y_{ij} = 1)$ since the data are binary - will be introduced. The probit function, denoted as Φ, is chosen for specifying the model examples for the comparative study. Consider

- *a) the generalized linear model* (regardless of the correlation structure between the observations of each subject) given by

$$\mathbb{P}(Y_{ij} = 1) = \Phi\left(\mu + \beta c_i\right),$$

- *b) the mixed effects generalized linear model* such that

$$\mathbb{P}(Y_{ij} = 1 | \eta_i) = \Phi\left(\mu + (\beta + \eta_i) c_i\right) \ , \ \eta_i \underset{i.i.d.}{\sim} \mathcal{N}(0, \omega^2),$$

- and *c) the Markov model* given by

$$\mathbb{P}(Y_{ij} = 1 | Y_{i,j-1} = k) = \Phi\left(\mu + \beta c_i + \alpha k\right).$$

In the last two examples, the marginal distribution of Y_{ij} can be derived, although not necessarily explicitly. Moreover, it generally does not lead to an equality with $\Phi\left(\mu + \beta c_i\right)$ as in the simple GLM case. In the mixed-effects GLM, it requires integration over the random effect η_i distribution:

$$\begin{aligned}\mathbb{P}(Y_{ij} = 1) &= \int \Phi\left(\mu + (\beta + \eta_i) c_i\right) p(\eta_i | \omega^2) d\eta_i, \\ &\neq \Phi\left(\mu + \beta c_i\right),\end{aligned}$$

whereas in the Markov model, it requires summing over all possible values for the past observations from time 0 up to time $j-1$:

$$\begin{aligned}\mathbb{P}(Y_{ij} = 1) &= \sum_{k_0=0}^{1} \sum_{k_1=0}^{1} \cdots \sum_{k_{j-1}=0}^{1} \left[\mathbb{P}(Y_{i0} = k_0) \prod_{l=1}^{j} \mathbb{P}(Y_{il} = k_l | Y_{i,l-1} = k_{l-1}), \right] \\ &\neq \Phi\left(\mu + \beta c_i\right),\end{aligned}$$

with $\mathbb{P}(Y_{il} = k_l | Y_{i,l-1} = k_{l-1})$ given by $\Phi\left(\mu + \beta c_i + \alpha k_{l-1}\right)$ if $k_l = 1$ and $1 - \Phi\left(\mu + \beta c_i + \alpha k_{l-1}\right)$ otherwise.

Figure 2 illustrates how $\mathbb{P}(Y_{ij} = 1)$ changes according to the value of β in cases *a)*, *b)* and *c)*. The larger ω, the more distant the marginal distributions of *b)* and *a)* will be. Moreover, while the marginal distribution remains common to all the Y_{ij}'s of a given subject with fixed and random effects GLMs, it does differ from time point to time point in the Markov model. This corroborates the fact that the interpretation of parameter β is completely different according to the model. In the fixed-effects GLM, β characterizes the effect of covariate c_i on the probability for observing presence $Y_{i,j} = 1$ at time j at the population level, whereas in the random-effects GLM, it has a similar meaning at the individual level, *i.e.* it is conditional to the individual parameters. Population and individual conclusions are of course combined when the between-subject variability is null. On the contrary, in model *c)*, β is to be interpreted in terms of transitions.

3 Simulation study

3.1 Objectives and simulation settings

The simulation study is performed to compare several longitudinal models with a specific focus on their predictive characteristics, on a whole cohort of patients (first simulation study) as well as on small groups of patients treated with similar dose levels (second simulation study). In both cases, the robustness to model misspecification, thanks to different simulation scenarios specified thereafter, is also explored. The whole study is based on original datasets of two similar Phase I studies on an MTA where the toxicity is measured on scale from 0 to 5. Since the models, data generation procedures and predictive criteria are common to the different steps of the simulation study, they are described once for clarity before summarizing the results of the two simulation studies.

3.1.1 Model construction from real clinical data

The subsequent simulation studies are based on four models which correspond each to one modelling approach introduced in Section 2. Each one assumes the subjects to be independent of one another. Two of these four models are Generalized Linear Models, with independent and correlated data respectively, and the others two are transitional models, with fixed and random effects respectively. These four models are built using two real datasets from similar early clinical trials investigating the same MTA. In these trials, the toxicity observations are binary outcomes ($K = 1$): $Y_{ij} = 0$ if the observed toxicity level is 0 or 1, $Y_{ij} = 1$ otherwise, and the observations are collected at regular times, *ie* $t_{ij} = j$ for all $i = 1, \ldots, N$ and all $j = 1, \ldots, n_i$. Many covariates are available to predict the toxicity responses. Some of these covariates are highly correlated as, for example, the dose level (*ie* the dose administered at each administration) and the dose intensity (*i.e.* the average administered dose according to time of treatment). Since highly correlated covariates may provide poor parameter estimations and lead to possible misinterpretations, the available covariates are grouped into 4 categories and final covariate selection for each model is performed by Information criteria (AIC) under restrictions that no more than one covariate from each group may be included in the model. The following covariate categories are considered: -1. Individual characteristics like age or gender -2. Treatment period and time from the last drug administration (TFLDA) -3. Measure of administered treatment like dose level (DL) or dose intensity (DI) -4. Cumulated treatment exposure like cumulated dose (CumDose) or cumulated AUC of patients' pharmacokinetic curves (CumExp). For most of these covariates, the logarithmic transformations have also been proposed (and assigned to their corresponding categories). The following models are selected this way, with Φ denoting for the probit link:

- **GLM (Generalized linear model)**: for each subject i, the temporal toxicity observations are given by:

  $$\mathbb{P}(Y_{ij} = 1|\varphi_0) = \Phi(\mu^{(0)} + \beta_1^{(0)} \log t_{ij} + \beta_2^{(0)}\text{TFLDA}_{ij} + \beta_3^{(0)} \log \text{DI}_i + \beta_4^{(0)}\text{CumDose}_{ij}),$$

  $j = 1, \ldots, n_i$, where $\varphi_0 = (\mu^{(0)}, \beta_1^{(0)}, \beta_2^{(0)}, \beta_3^{(0)}, \beta_4^{(0)})$.

- **AR (Marginal model with autoregressive correlation structure)**: for each subject i, the moments of the temporal toxicity observations are assumed to be such that:

  $$\begin{aligned}\mathbb{P}(Y_{ij}=1|\varphi_1) &= \Phi(\mu^{(1)}+\beta_1^{(1)}t_{ij}+\beta_2^{(1)}\mathrm{TFLDA}_{ij}+\beta_3^{(1)}\mathrm{DI}_i+\beta_4^{(1)}\log\mathrm{CumExp}_{ij}),\\ \mathbf{Corr}(Y_{i,j},Y_{i,k}|\varphi_1) &= \rho^{|k-j|},\end{aligned}$$

  $1\leq j,k\leq n_i$, where $\varphi_1=(\mu^{(1)},\beta_1^{(1)},\beta_2^{(1)},\beta_3^{(1)},\beta_4^{(1)},\rho)$.

- **F.markov (Fixed-effects Markov model)**: for each subject i, the temporal toxicity observations are given by:

  $$\mathbb{P}(Y_{ij}=1|Y_{i,j-1},\varphi_2)=\Phi(\mu^{(2)}+\beta_1^{(2)}t_{ij}^2+\beta_2^{(2)}\log\mathrm{DI}_i+\beta_3^{(2)}\mathrm{CumExp}_{ij}+\alpha^{(2)}Y_{i,j-1}),$$

  where $\varphi_2=(\mu^{(2)},\beta_1^{(2)},\beta_2^{(2)},\beta_3^{(2)},\alpha^{(2)})$.

- **M.markov (Mixed-effects Markov model)**: given some subject-specific parameters φ_i, the temporal observations of each subject are given by the same structural model:

  $$\mathbb{P}(Y_{ij}=1|Y_{i,j-1},\varphi_i)=\Phi(\mu_i^{(3)}+\beta_{i1}^{(3)}\log t_{ij}+\beta_{i2}^{(3)}\mathrm{TFLDA}_{ij}+\beta_{i3}^{(3)}\mathrm{DoseLevel}_i+\alpha_{i1}^{(3)}Y_{i,j-1}).$$

  Here, the individual parameters are $\varphi_i=(\mu_i^{(3)},\beta_{i1}^{(3)},\beta_{i2}^{(3)},\beta_{i3}^{(3)},\beta_{i4}^{(3)},\alpha_{i1}^{(3)})$. To model the between-patient variability, the simplifying assumption that only parameter $\alpha_{i1}^{(3)}$ is a random effect: $\alpha_{i1}^{(3)}\underset{i.i.d.}{\sim}\mathcal{N}(\alpha_1^{(3)},\omega_\alpha^2)$, whilst the other individual parameters are shared among subjects with ignorable variation such that $\forall i\mu_i^{(3)}\equiv\mu^{(3)}$, $\beta_{i\ell}^{(3)}\equiv\beta_\ell^{(3)}$, $\ell=1,\ldots,4$. Thus, the population parameters are $\theta=(\mu^{(3)},(\beta_\ell^{(3)})_{\ell=1,\ldots,4},\alpha_1^{(3)},\omega_\alpha^2)$.

Parameter estimation in **GLM** and **F.markov** is provided by maximization of the likelihood (Iteratively Re-weighted Least Squares algorithm [21]), **AR** parameters estimation is based on a Penalized Quasi-Likelihood maximization method [25, 24] and in **M.markov**, it is given by maximization of the likelihood through the Gauss-Hermite quadrature method [27].

3.1.2 Predictive criteria

The ensemble (*i.e.* whatever the dose level) predictive performance and the dose-specific predictive performance of models **GLM**, **AR**, **F.markov** and **M.markov** will be assessed according to the three criteria below.

1. First, the ability for each model to predict which patients may or may not have a toxicity event over their period of participation in the trial will be examined. This will be evaluated by means of the Brier score [28], which is a proper score generally used to quantify the quality of such binary predictions. This score reaches its minimum when a perfect matching between the probabilistic forecast and the distribution of the event to be predicted is obtained.

2. The second examined criterion is how accurately each model predict the number of adverse events occurring for each patient during the trial. Since the number of reported toxicities induced by MTAs is likely to be larger for long periods than for short ones, the number of adverse events is scaled according to the duration of the follow-up. The predictions of the number of adverse events are then evaluated with the Continuous Rank Probability Score (CRPS) [29], which is also a proper score. Thus, the better the model predicts the number of adverse events, the smaller the CRPS will be.

3. Third, the predicted number of weeks with a toxicity response is controlled. As for previous criteria, the number of weeks with a toxicity response is brought back to the duration of the follow-up and the CRPS is used to evaluate the ability for each model to provide such good predictions.

3.2 Simulation study 1: comparison on overall predictive characteristics

This first simulation study aims at comparing the predictive properties of models **GLM**, **AR**, **F.markov** and **M.markov** without any distinction between the patient dose levels of the MTA of interest.

3.2.1 Data generation

Three simulation scenarios are considered. To simulate realistic data, individual profiles (dose level, drug exposure, time of treatment, etc) are sampled with replacement from the original datasets. Under *Scenario1*, a sort of bootstrap scenario, the patients' profiles (*i.e.* values for the covariates and length of follow-up) and the corresponding toxicity longitudinal responses are sampled with replacement from the original dataset. Such a simulation technique will help identify the best predictive model without any idea of the true model. Two other simulation scenarios, *Scenario2* and *Scenario3*, are then used to investigate the robustness to model misspecification. In both, the simulated patients' clinical characteristics are obtained by random sampling with replacement from the initial data as in *Scenario1*, but the corresponding toxicity longitudinal responses are simulated according to mixed-effects Markov model **M.markov** (using the parameters values $\mu^{(3)} = -6.85$, $\beta_1^{(3)} = 1.30$, $\beta_2^{(3)} = 0.14$, $\beta_3^{(3)} = 0.01$, $\alpha_1^{(3)} = 1.53$ and $\omega_\alpha = 0.48$) and marginal model with autoregressive correlation structure **AR** respectively (using the parameters values $\mu^{(1)} =$, $\beta_1^{(1)} =$, $\beta_2^{(1)} =$, $\beta_3^{(1)} =$, $\beta_4^{(1)} =$ and $\rho =$). Since marginal models are rarely completely specified, it is hard to simulate data according to such models. Thus, the simulation technique introduced in [30] is used to generate the data under *Scenario3*.

For each of these three data generation scenarios, the procedure is the following. 1000 pairs of independent datasets with N patients each are simulated. The first one (Tr_k) is a "training sample" and the second one (Te_k) is a "test sample", $k = 1, \dots, 1000$. For each value of k, the four models **GLM**, **AR**, **F.markov** and **M.markov** are inferred from Tr_k regardless of the scenario used for generating the data, whereas the predictive properties of the models are computed by Monte Carlo from Te_k by comparing each patient's responses from Te_k with 1000 probable longitudinal toxicity responses under the estimated model. Two sample sizes are investigated: $N = 30$, which corresponds more or less to the size of most phase I studies, and $N = 50$, which is the size of the original datasets used to design the present study.

3.2.2 Results

Figures 3, 4 and 5 display the results related to the prediction of at least one toxic event, to the prediction of the duration of adverse events and to the prediction of the total number of adverse events respectively. On each figure, the first line displays the model performance under the three simulation scenarios in cases of small datasets ($N = 30$) and the second line displays the model performance under the three simulation scenarios in cases of larger datasets ($N = 50$).

The first noticeable result is that under the bootstrap simulation scenario (*Scenario 1*) and whatever the sample size, models **AR** and **M.markov** lead to Brier scores and CRPSs smaller than those given by models **GLM** and **F.markov**, without much difference between the performance of **AR** and **M.markov**. This especially means that the assumption on the nature of the dependence between longitudinal toxicity observations underlying these two models provides an adequate description of the toxic reactions.

Simulation scenarios *Scenario 2* and *Scenario 3* are aimed at evaluating the robustness of the candidate models to model misspecification. Under *Scenario 2*, data are generated according a mixed-effects Markov model. Figures 3, 4 and 5 show that the best predictive model according to the three chosen criteria is either the fixed-effects Markov model (**F.markov**) or the mixed-effects Markov model (**M.markov**) when $N = 30$ and always the mixed-effects Markov model when $N = 50$. The difficulty of recovering the true model from small datasets is probably linked to the difficulty of correctly evaluating the between-patient variability in this case but the main assumption of a Markov structure of dependence between observations is sufficient to provide correct predictions. Under *Scenario 3*, data are generated according to a marginal model with an autoregressive correlation structure. Under this scenario, no clear difference is evidenced between the four models whatever the sample size. Indeed, the median and the dispersion of the three scores are quite similar from one model to another.

As an overall conclusion, this first simulation study shows that assuming a dependence structure (autoregressive or Markov) between longitudinal toxicity records leads to the most appropriate predictions of the toxic reactions to MTAs. More precisely, Markov models, in particular mixed effects Markov models, would be a robust model choice for accurately describing longitudinal toxicity records.

3.3 Simulation study 2: comparison on dose-specific predictions

Beyond the comparison of models **AR** and **M.markov** according to their overall performances in predicting toxic events for entire cohorts of patients, their predictive performance is also compared on the scale of subgroups of patients having been given the same dose levels. This second simulation study aims at assessing the ability of each model to discriminate between doses, thus at assessing the appropriateness of each model in dose-finding.

3.3.1 Data generation

Two situations are investigated: *a)* an ideal case where the dose groups are balanced, and *b)* a more realistic case where the patients' allocation to dose groups is not balanced. 5 dose groups are considered, further denoted as DI_1, DI_2, DI_3, DI_4 and DI_5. In situation *b)*, the number of simulated patients in each dose group varies from one simulated dataset to another, since in this case, dose level is assigned by randomly choosing a patient and his dose level among those of the original datasets. In both balanced and unbalanced configurations, 1000 pairs of cohorts of 50 patients are simulated assuming the same monitoring duration for all patients, including 9 weeks of treatment and 6 weeks of follow-up. The patients' drug exposure (*i.e.* the values over time of explanatory variable CumExp) is simulated according to a log-linear model, previously inferred from the original datasets: $\log CumExp_{ij} = \mu + u_i + \beta_1 \times t_{ij} + \beta_2 \times \log(t_{ij}) + \beta_3 \times DI_i + beta_4 \times \log(DI_i) + \varepsilon_{ij}$, where $\mu = -16.4$, $u_i \sim \mathcal{N}(0, 1.75^2)$, $\beta_1 = -0.011$, $\beta_2 = 1.02$, $\beta_3 = -0.027$, $\beta_4 = 4.96$ and $\varepsilon_{ij} \sim \mathcal{N}(0, 0.16^2)$. Their dose intensities are also simulated. As in previous simulation study, robustness of predictions by dose level to model misspecification is assessed by simulating the patients' toxicity events according to different scenarios. Under *Scenario 4*, a mixed effects Markov model (**M.markov**) is used to simulate the toxicities, with the following parameter values: $\mu^{(3)} = -6.3, \alpha_1^{(3)} = 0.9, \beta_1^{(3)} = 0.9, \beta_2^{(3)} = -0.1, \beta_3^{(3)} = 0.025$, $\omega_\alpha^2 = 1$; and under *Scenario 5* the toxicities are simulated according to a marginal model with autoregressive correlation structure (**AR**), using parameter values $\mu^{(1)} = -1.78, \beta_1^{(1)} = 0.156, \beta_2^{(1)} = -0.128, \beta_3^{(1)} = 0.01, \beta_4^{(1)} = 0.153, \rho = 0.83$. The corresponding theoretical probabilities of having at least one toxic event over the 15 weeks of monitoring are 0.06, 0.11, 0.19, 0.3, 0.41 respectively for doses DI_1, DI_2, DI_3, DI_4 and DI_5 under *Scenario 4* and 0.08, 0.19, 0.32, 0.42, 0.50 under *Scenario 5*.

Each pair of simulated datasets includes, as above, one training set and one validation set. Whatever the scenario used for data simulation (*Scenario 4* or *Scenario 5*), the procedure is the following. For each of the 1000 simulated training cohorts, the two competing models (**AR** and **M.markov**) are first inferred; then, using the estimated model parameters, the predicted probability of having at least one toxic event is computed for each patient of the corresponding validation cohort. The empirical distributions of the predicted probabilities of having at least one toxic event over the 15 weeks of monitoring can be deduced from the 1000 simulated cohorts. It is then compared to the theoretical probabilities stated above using the CRPS.

3.3.2 Results

The results are depicted Figure 6. The first line gives the results obtained when simulating balanced dose level groups, whereas the second line displays the results arising from simulating unbalanced dose level groups. On the left, data are simulated according to a mixed-effects hidden Markov model (*Scenario 4*); on the right, data are simulated according to a generalized linear model with an autoregressive correlation structure (*Scenario 5*). In case *a)* (balanced dose level groups), model **AR** provides better predictions on low dose intensities (DI_1 and DI_2) whatever the scenario used for data simulation, while model **M.markov** tends to provide better predictions on higher dose intensities (from DI_3 to DI_5), which are closer to the MTD (e.g. 35%). In case *b)* (unbalanced dose level groups), model **M.markov** provides better overall predictions whatever the scenario used for data simulation. Indeed, the CRPS distributions for each dose group show a lower median score and smaller variability in prediction when obtained by inferring the autoregressive model.

4 Discussion

4.1 Categorical and temporal extensions

One can imagine many extensions of the longitudinal data models described above. First of all, although the illustrative examples presented in Sections 2.3 and 3.1.1 restrict for simplicity the temporal toxicity responses to binary outcomes (typically presence or absence of side effects), we emphasize that the approach also applies to more general types of outcomes like ordinal categorical ones. The model can consider different grades of toxicity, for instance from 0 to 5 according to NCI-CTCAE nomenclature, which would allow for a more sophisticated description of the evolution of toxicity over time with respect to the clinician's interest. Secondly, our presentation focused on modelling some unidimensional temporal toxicity responses. However, just like standard oncology therapies, MTAs are likely to induce several kinds of toxic effects like, without being exhaustive, ocular toxicities, peripheral neuropathies or urinary toxicities. When determining the Maximum Tolerated Dose, one should take into account the way all these toxicities jointly occur over time, for which the unidimensional longitudinal models presented in Section 2 offer limited prospects. There is indeed no reason to consider that the underlying mechanisms for these different categories of toxicities are the same or that these toxic effects arise independently. Therefore, it is of strong interest to ultimately extend the models above to (possibly correlated) multidimensional temporal responses. Data augmentation as in [31] is the convenient entry point to develop hierarchical stochastic structures with increased complexity: introducing a latent multivariate (normal autoregressive) layer is the price to pay in order to match all these operational requirements.

By nature, the serial dependence is implemented in the Markov models we considered, in addition to a between-individual variability induced by the random effects for the last one. A hidden Markov model (HMM) may be an attractive alternative for the case study ([32]). Instead of establishing a direct transition between observable states Y_{ik} as given by equation 2, in HMM, the time dependence stems from a latent Markov model Z. For instance a first order Normal AR model would be written:

$$\begin{aligned} &Z_{ij} = \rho_i Z_{i\,(j-1)} + \sigma_i \varepsilon_{ij} \\ &\varepsilon_{ij} \sim N(0,1). \end{aligned}$$

Although ρ_i can be high to keep the latent process Z_{ij} tight enough during long sequences, some dispersion is nevertheless allowed thanks to the stochastic observation equation given by:

$$\mathbb{P}(Y_{ij} = 1 | Z_{ij}) = \Phi\left(\mu + \beta c_i + Z_{ij}\right)$$

Such HMM with random effects ($(\rho_i, \sigma_i) \sim p(.|\theta)$) have recently been proposed by [33] in the advanced statistical literature.

4.2 On the path of personalized medicine

The present study mainly focused on the interests of longitudinal data models at the population scale, the idea being to depict the mechanism of the appearance of toxic effects over time and their variations around some population average. But the other aspect of interest for longitudinal data models lies at the individual level. Once the model is inferred from the whole trial data, one can take advantage of its predictions at the individual levels. Borrowing strength from neighbours but allowing for some heterogeneity among individuals, random effect models seem here to be efficient tools purposely tailored.

4.3 Missingness

Getting personalized knowledge about individuals appears particularly relevant in the context of phase I clinical trials in oncology since overly strong adverse events lead to early withdrawal from the trial. It is worth noting that such early drop-out cannot be considered independently of the observable and missing data. Describing the missingness mechanism yields additional modelling and inference challenges not dealt with this paper as well as extra programming not available in standard software. [20] provides an overview of the various modelling frameworks for non-Gaussian longitudinal data, with non ignorable missingness processes. Once a patient shows a critical toxicity profile, the model can help predict his near future and anticipate stronger adverse effects and can therefore possibly adapting his treatment for more comfort.

4.4 Words of caution

The introduction of longitudinal models in contexts of phase I clinical trials in oncology however is not exempt from some caveats. First, longitudinal models allow the use of time-varying covariates to predict the toxicity outcomes, which was not possible with approaches based on a one-shot outcome (non longitudinal models). Clearly, these time-varying covariates are a very rich source of information but also a potential source of confusion with the temporal memory of toxicity responses. It is however up to the user to use these time-varying covariates cautiously in a way to avoid misinterpretation of the model parameter estimations. A typical situation where some problems may arise is when the treatment is readjusted (delayed administration or reduced dose administration) because of the occurrence of disabling or persistent toxic effects. In such a situation, the causal relationship between toxicities and drug exposure becomes ambiguous since it works in both directions.

A second warning concerns the amount of data required to satisfactorily infer complex longitudinal models. The more sophisticated the model becomes, the more data is needed to get trustworthy parameter estimations. Yet the size of studied groups in phase I clinical trials is often rather limited, from say 20 to 80 patients. As usual, model complexity has to be weighted with sample size.

5 Conclusion

5.1 From retrospective analysis to experimental design

In this paper, the retrospective analyses based on data collected from phase I clinical trials in oncology plead in favour of the use of longitudinal models, especially Markovian models with random effects. We believe that these advanced stochastic structures are not limited to the statistical description of the data, but in this section we show that they could also develop as operational tools. Our work was particularly motivated by the recent development of MTAs, which rely on more complex mechanisms of action than do standard oncology treatments. Multiple dose administrations to the same patient are performed over multiple cycles. MTAs provoke repeated side effects of variable intensity in the long term that are recorded in the follow-up of each patient (for instance such follow-up may last over six cycles of one month). In the practice of clinical trials however, the next cohort of patients to be treated is launched long before all the data of the previous cohort has been completely recorded, say on a monthly basis. To cope with this short horizon, most studies currently reduce the data from the previous cohorts to a binary end point, the occurrence of a toxicity, ignoring the incompleteness of the data, the recurrence pattern of events and the various grades of toxicity. Such preliminary "syntheses" to sum up the follow-up data yield a substantial loss of information, particularly on the temporal evolution of toxicities, and therefore present the risk of recommending an unadapted dose. Longitudinal data models give clinicians the opportunity to take advantage of all available information from the clinical trial. Following the same avenue of thought of the continuous reassessment method, more refined dose-finding strategies could be elaborated after updating the knowledge about the longitudinal model parameters as information accumulates and late toxicities from previous cohorts come out.

5.2 Avoiding dose limiting toxicities

Up to now, unless a dose limiting toxicity occurs, the dose received by a patient is usually not changed from one cycle to the next. Longitudinal models intend to assess the mechanisms underlying the appearance of toxicities over time and allow for the incorporation of time-varying covariates. They account for the dependence between consecutive observations and thus make us able to assess different toxicity profiles like late, repeated or persistent toxicities. We show that these models also have interesting predictive properties for separate pre-defined dose categories. Therefore the extra information gained from each cycle provides more precise estimation to understand the individual patient's behaviour with respect to toxicity during the follow-up period and might even enable a better adaptation of the dose within the cycles of a given patient.

5.3 Redefining the MTD

However over-simplistic the statistical approaches presently retained for phase I clinical trials in oncology may be, they have the great merit of designing clinical trials targeting MTD. Nevertheless, as a direct end-product of these simple models, the MTD is in essence a statistical quantile. The single summarizing outcome per patient follow-up is considered as a Bernoulli event, and the MTD is merely the dose corresponding to the 33% quantile of the probability model that rules such Bernoulli events, with one shot per patient. Now, phase I oncology experts have come to the agreement that new recommendations for toxicity evaluation are needed, in particular requiring the assessment of new MTD based on multiple endpoints [8, 34].

In the present work, we study alternative models allowing us to learn from multivariate longitudinal categorical data. Unlike previous one-shot Bernoulli models, such longitudinal models offer the appropriate framework to enrich the definition of the Maximum Tolerated Dose (which remains the main objective of a Phase I clinical trial in oncology) by taking into consideration different endpoints longitudinally followed. A renewed definition of the Maximum Tolerated Dose appropriate to MTA compounds should take into consideration acute, late-onset, repeated or prolonged toxicities of many kinds along with their related severity. For instance, the MTD could be seen as the maximum dose verifying that different toxic events occur with some predefined controlled risks:

$$\arg_{d}\left(\Big(\mathbb{P}(E_1|d) < t_1\Big) \wedge \Big(\mathbb{P}(E_2|d) < t_2\Big) \wedge, ..., \wedge \Big(\mathbb{P}(E_S|d) < t_S\Big)\right)$$

where $E_s (s = 1 \dots S)$ are the toxic events of interest and t_s the consented risks of occurrence of these events. As an example, E_s could be a late onset peripheral neuropathy event with a grade above or equal to 3 according to NCI-CTCAE, or a Grade 2 to 4 ocular toxicity event lasting more than 4 weeks, or even the simultaneity of those two toxic events. Inasmuch as enough data becomes available to perform a relevant inference, the proposed longitudinal data models allow MTD to be assessed based on the richer definitions that phase I oncologists are willing to promote.

Acknowledgements The authors thank Joan Sobota for her advice and comments.

References

[1] Booth CM, Calvert AH, Giaccone G, Lobbezoo MW *et al.* On behalf of the Task Force on Methodology for the Development of Innovative Cancer Therapies, Endpoints and other considerations in phase I studies of targeted anticancer therapy: Recommendations

from the task force on Methodology for the Development of Innovative Cancer Therapies (MDICT). *European Journal of Cancer*. 2008; **44**:19-24. DOI: 10.1016/j.ejca.2007.07.034.

[2] Ratain MJ, Humphrey RW, Gordon GB, Fyfe G *et al.* Recommended changes to oncology clinical trial design: revolution or evolution? *European Journal of Cancer*. 2008;**44**:8-11. DOI: 10.1016/j.ejca.2007.09.011

[3] Arrowsmith J. Phase II failures: 2008–2010. *Drug Discovery*. 2011;**10**:328-329. DOI: 10.1038/nrd3439

[4] Ji Y, Liu P, Li Y, Bekele N. A modified toxicity probability interval method for dose-finding trials. *Clinical Trials*. 2010;**7**:653-663. DOI: 10.1177/1740774510382799

[5] O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*. 1990;**46**:33-48. DOI: 10.1191/1740774506cn134oa

[6] Babb J, Rogatko A, Zacks S. Cancer Phase I clinical trials: Efficient dose escalation with overdose control. *Statistics in Medicine*. 1998;**17**: 1103-1120. DOI: 10.1002/(SICI)1097-0258(19980530)17:10<1103::AID-SIM793>3.0.CO;2-9

[7] Neuenschwander B, Branson M, Gsponer T. Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in Medicine*. 2008;**27**: 2420-2439. DOI: 10.1002/sim.3230

[8] Soria JC. Phase 1 trials of molecular targeted therapies : Are we evaluating toxicities properly? *European Journal of Cancer*. 2011;**47**:1443-1445. DOI:10.1016/j.ejca.2011.04.009

[9] Postel-Vinay S, Gomez-Roca C, Molife R, Anghan B *et al.* Phase I Trials of Molecularly Targeted Agents: Should We Pay More Attention to Late Toxicities? *Journal of Clinical Oncology*. 2011;**29**:1728-1735. DOI: 10.1200/JCO.2010.31.9236

[10] Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*. 2000;**56**:1177-82. DOI: DOI: 10.1111/j.0006-341X.2000.01177.x

[11] Di Scala L, Pylvänäinen I, Molloy B, Manlius C *et al.* A Novel Bayesian Dose-Escalation Phase Ib Design Investigating Safety of Combination of RAD001 (Everolimus) With Chemotherapy Plus Trastuzumab in Patients With HER2-Overexpressing Metastatic Breast Cancer With Prior Resistance to Trastuzumab. *Journal of Clinical Oncology*, 2008; ASCO Annual Meeting Proceedings (Post-Meeting Edition). **26**(15S):1130

[12] Rogatko A, Babb JS, Wang H, Slifker MJ *et al.* Patient characteristics compete with dose as predictors of acute treatment toxicity in early phase clinical trials. *Clinical Cancer Research.* 2004;**10**:4645-51. DOI:10.1158/1078-0432.CCR-03-0535

[13] Chen Z, Krailo MD, Azen SP, Tighiouart M. *A novel toxicity scoring system treating toxicity response as a quasi-continuous variable in Phase I clinical trials.* Contemporary Clinical Trials, 2010, 31(5), 473-482. DOI:10.1016/j.cct.2010.05.010

[14] Lee SM, Hershman DL, Martin P, Leonard JP *et al.* Toxicity burden score: a novel approach to summarize multiple toxic effects. *Annals of Oncology.* 2011;**23**:537-541. DOI:10.1093/annonc/mdr146

[15] Lee SM, Cheng B, Cheung YK. Continual reassessment method with multiple toxicity constraints. *Biostatistics.* 2011;**12**:386-398. DOI:10.1093/biostatistics/kxq062

[16] Potthoff RF, George SL. Flexible Phase I Clinical Trials: Allowing for Nonbinary Toxicity Response and Removal of Other Common Limitations. *Statistics in Biopharmacy Research.* 2009;**1**:213-228. DOI:10.1198/sbr.2009.0014.

[17] Bekele BN, Li Y, Ji Y. Risk-Group-Specific Dose Finding Based on an Average Toxicity Score. *Biometrics.* 2010;**66**:541-548. DOI:10.1111/j.1541-0420.2009.01297.x

[18] Doussau A, Thiébaut R, Paoletti X. Dose-finding design using mixed-effect proportional odds model for longitudinal graded toxicity data in phase I oncology clinical trials. *Statistics in Medicine.* 2013;**32**:5430-5447. DOI: 10.1002/sim.5960

[19] Diggle PJ, Liang KY, Zeger SL. Analysis of Longitudinal Data. *Oxford University Press*, 1994.

[20] Jansen I, Beunckens C, Molenberghs G, Verbeke G *et al.* Analyzing Incomplete Discrete Longitudinal Clinical Trial Data. Statistical Science. *Statistical Sciences.* 2006;**21**:52-69. DOI: 10.1214/088342305000000322

[21] McCullagh P, Nelder J. Generalized Linear Models, Second Edition, *Taylor & Francis*, 1989 - 532 pages.

[22] Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics.* 1986;**42**:121-30.

[23] Xu H, Craig BA. Likelihood Analysis of Multivariate Probit Models Using a Parameter Expanded MCEM Algorithm. *Technometrics.* 2010;**52**:320-348. DOI: 10.1198/TECH.2010.09055

[24] Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association.* 1993;**88**:9-25. DOI: 10.1080/01621459.1993.10594284

[25] Wolfinger R, O'Connell M. Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation.* 1993;**48**:233-243. DOI: 10.1080/00949659308811554

[26] Davidian M, Giltinan DM. Nonlinear Models for Repeated Measurement Data. *Chapman and Hall/CRC*; 1 edition, 1995.

[27] Pinheiro JC, Bates DM. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics.* 1995;**4**:12-35.

[28] Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review.* 1950;**78**:1-3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[29] Epstein ES. A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology.* 1969;**8**:985-987. DOI: 10.1175/1520-0450(1969)008<0985%3AASSFPF>2.0.CO%3B2

[30] Biswas A. Generating correlated ordinal categorical random samples. *Statistics & Probability Letters.* 2004;**70**:25-35. DOI:10.1016/j.spl.2004.08.001

[31] Albert JH, Chib S. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association.* 1993;**88**:669-679. DOI:10.1080/01621459.1993.10476321

[32] Cappe O, Moulines E, Ryden T. Inference in Hidden Markov Models. 2005 *Springer, New York.*

[33] Altman RJ. Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association.* 2007;**102**:201-210.

[34] Postel-Vinay S, Collette L, Paoletti X, Rizzo E *et al.* Towards new methods for the determination of dose limiting toxicities and the assessment of the recommended dose for further studies of molecularly targeted agents - Dose-Limiting Toxicity and Toxicity Assessment Recommendation Group for Early Trials of Targeted therapies, an European Organisation for Research and Treatment of Cancer-led study. *European Journal of Cancer.* 2014;**50**:2040-2049. DOI: 10.1016/j.ejca.2014.04.031

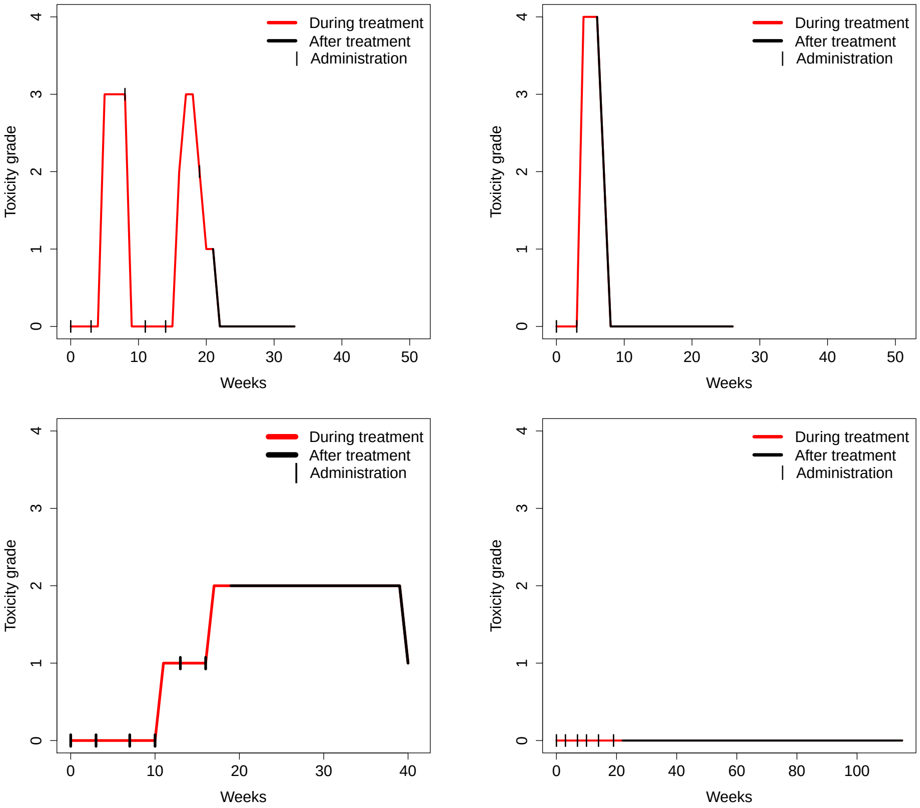

Figure 1: Typical toxicity records from an early phase clinical trial of a specific MTA. Each graph corresponds to a patient and represents the evolution over weeks of the intensity of toxic effects after drug intake on a scale from 0 (no toxic effect) to 4 (disabling effect).

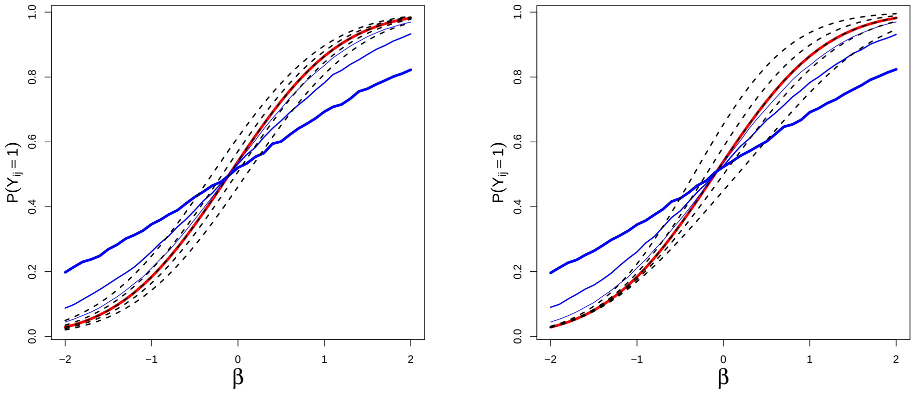

Figure 2: Relation between $\mathbb{P}(Y_{ij} = 1)$ and β in models *a)* (red), *b)* (blue) for different values of ω (the greater ω is, the more the thicker the line is) and *c)* (black dotted lines) for different values of α. Each graph corresponds to a different time point j ($j = 1$ on the left, $j = 2$ on the right). The following values are used: $\mu = 0.1$ and $c_i = 1$.

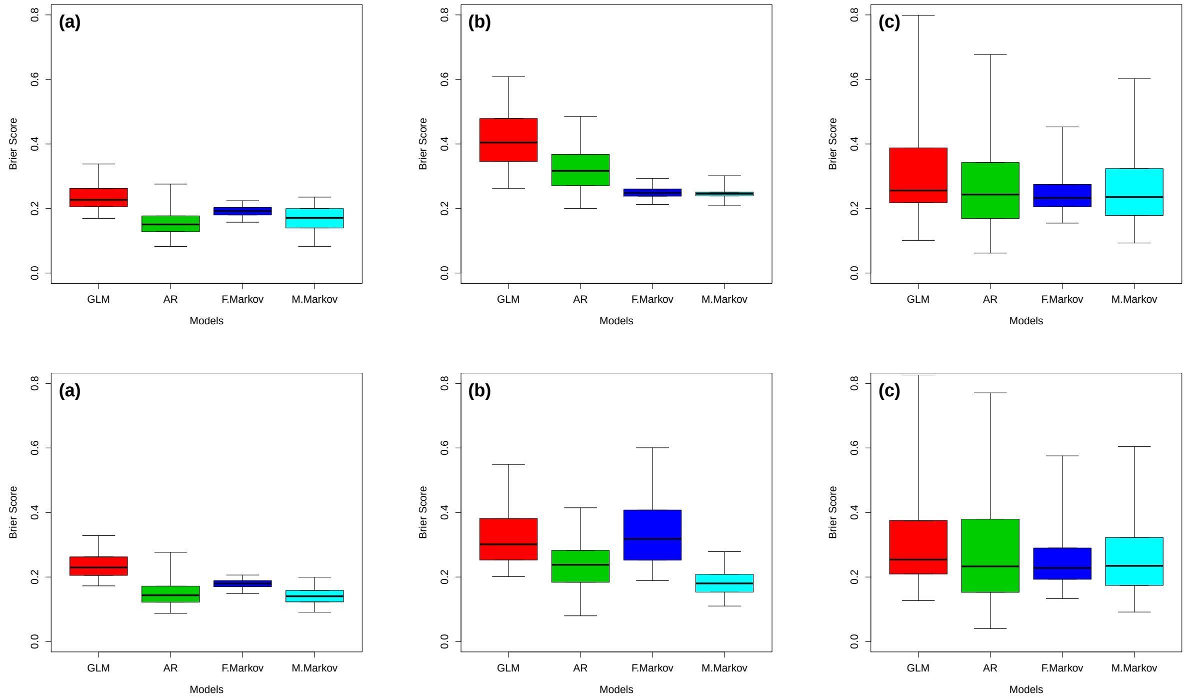

Figure 3: Brier score distributions obtained among 1000 simulations for each model and each simulation scenario. The first line and the second line respectively display the results for small datasets $N = 30$ and large datasets $N = 50$. From left to right on each line, the results obtained under Scenario1 (a), Scenario2 (b) and Scenario3 (c) are displayed. Whiskers represent the 95% confidence intervals. The Brier score is used to evaluate the model capability to predict the probability of having at least one toxic event. The lower the Brier score is, the better the model prediction is.

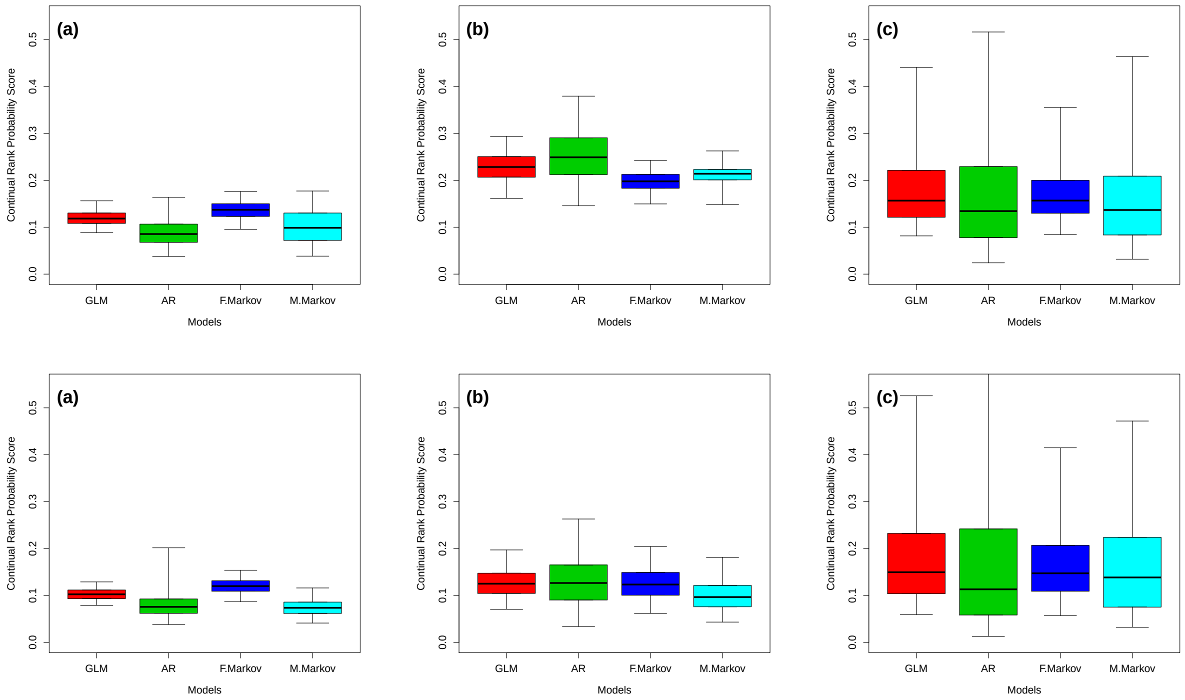

Figure 4: Continuous Rank Probability Score distributions for the number of weeks with adverse event obtained among 1000 simulations for each model and each simulation scenario. The first line and the second line respectively display the results for small datasets $N = 30$ and large datasets $N = 50$. From left to right on each line, the results obtained under Scenario1 (a), Scenario2 (b) and Scenario3 (c) are displayed. Whiskers represent the 95% confidence intervals. The lower the CRPS score is, the better the model prediction is.

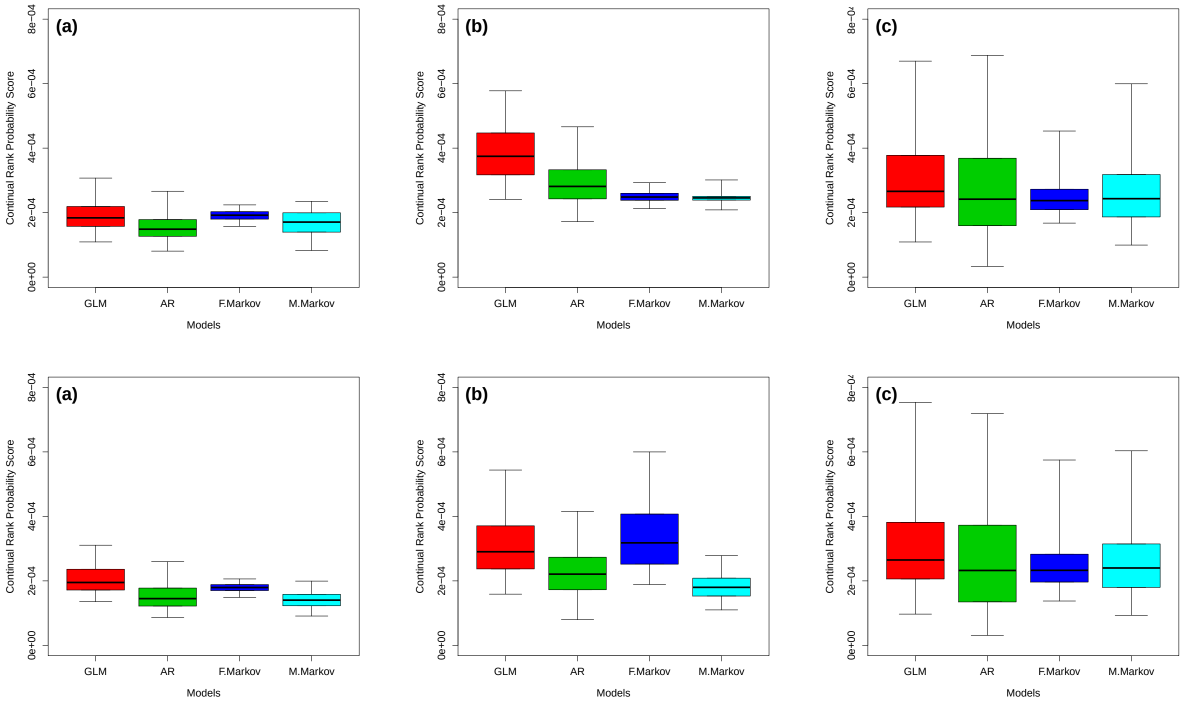

Figure 5: Continuous Rank Probability Score distributions for the number of Adverse Events obtained among 1000 simulations for each model and each simulation scenario. The first line and the second line respectively display the results for small datasets $N = 30$ and large datasets $N = 50$. From left to right on each line, the results obtained under Scenario1 (a), Scenario2 (b) and Scenario3 (c) are displayed. Whiskers represent the 95% confidence intervals. The lower the CRPS is, the better the model prediction is.

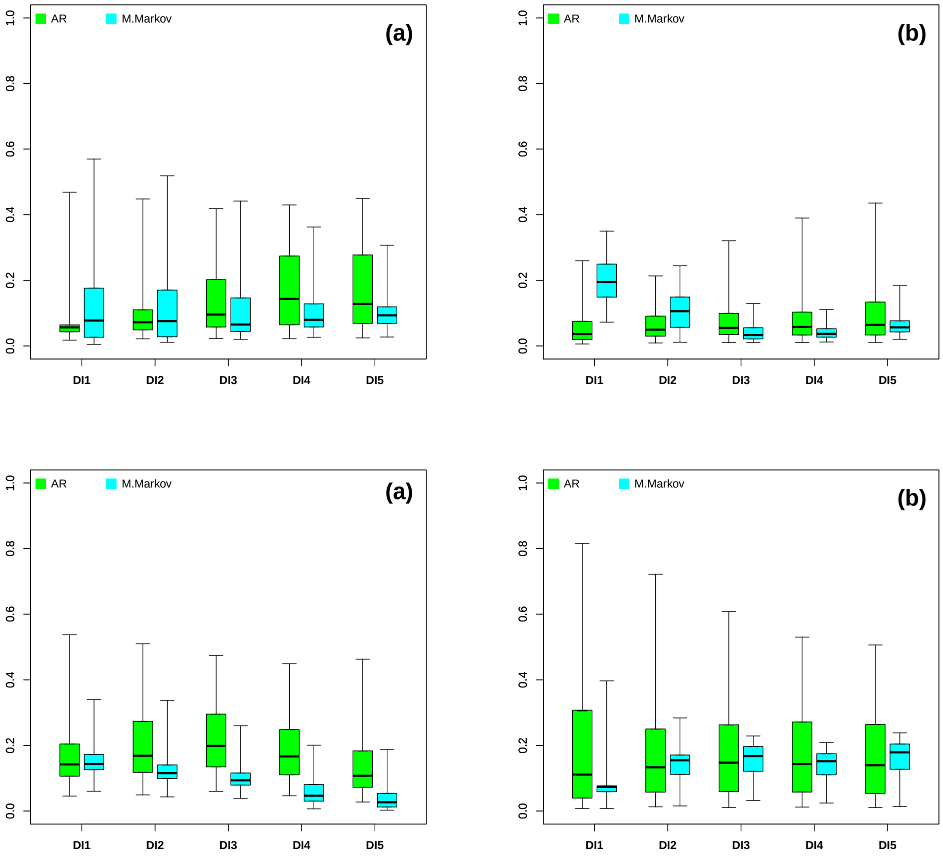

Figure 6: Continuous Rank Probability Score distributions applied on the predicted probability of having at least one adverse event over the 15 weeks of follow-up for several dose intensity values according to models **AR** and **M.markov**. Whiskers represent the 95% confidence intervals. The lower the CRPS is, the better the model prediction is. The first line depicts results obtained on 1000 cohorts of 50 patients allocated in balanced dose level groups (10 patients per dose level). The second line: depicts results obtained on 1000 cohorts of 50 patients allocated in unbalanced dose level groups. Left: longitudinal toxicity responses are simulated according to a mixed effect Markov model *Scenario 4*. Right: longitudinal toxicity responses are simulated according to a fixed-effects generalized linear model with autoregressive correlation structure *Scenario 5*.