Utility-based dose selection for phase II dose-finding studies

Jihane AOUNI[1,2], Jean Noel BACRO[2], Gwladys TOULEMONDE[2,3],
Pierre COLIN[1], and Loic DARCHY[1]

[1]Sanofi, Research and Development, 91385 Chilly-Mazarin, France
[2]IMAG, Univ Montpellier, CNRS, Montpellier, France
[3]Lemon, INRIA

November 18, 2019

Abstract

Dose selection is a key feature of clinical development. Poor dose selection has been recognized as a major driver of development failure in late phase. It usually involves both efficacy and safety criteria. The objective of this paper is to develop and implement a novel fully Bayesian statistical framework to optimize the dose selection process by maximizing the expected utility in phase III. The success probability is characterized by means of a utility function with two components, one for efficacy and one for safety. Each component refers to a dose-response model. Moreover, a sequential design (with futility and efficacy rules at the interim analysis) is compared to a fixed design in order to allow one to hasten the decision to perform the late phase study. Operating characteristics of this approach are extensively assessed by simulations under a wide range of dose-response scenarios.

Keywords: Bayesian approach; Dose selection; Interim analysis; Sequential trials; Utility function

1 Introduction

Until today, estimation of the right dose remains a key problem in drug development. It is now well documented that poor dose selection is a root cause for failures or delays in drug approval[1] (see also[2] guidelines). It is now also well accepted that finding the right dose should be rather considered as an estimation problem[3] than a multiple testing problem. This latter traditional approach, as well as the more recent Multiple Comparison Procedure and Modeling (MCP-Mod) methodology[4,5,6] generally consider efficacy and safety sequentially: doses associated with statistically significant differences versus the control, for the multiple testing approach, or doses with clinically relevant difference versus control, for the MCP-Mod approach, are identified first and then the highest dose amongst them considered as "well tolerated" is generally chosen. An alternative approach should rather rank the doses using efficacy and safety assessments simultaneously. On the other hand, in many settings the dose selection is mainly driven by efficacy only. In absence of safety considerations one typically searches for the dose which is near the plateau, e.g. the dose reaching 90% or 95% of the maximal efficacy denoted by ED90 and ED95. This holds for monotonic dose-responses. Higher doses will unduly expose the patients to potential toxicity issues while lower doses may represent a substantial loss of efficacy. Another dose of interest is the Minimum Effective Dose (MED), i.e. the smallest dose associated with a statistically significant and clinically relevant effect. The range of doses between the MED and the ED90/ED95 constitutes the interesting dose zone[7]. When serious safety issues arise within this interesting zone, the dose selection becomes more challenging and involves multiple criteria. Additional toxicity may counterbalance a gain of efficacy and one needs to introduce some utility score balancing both efficacy and safety.

This paper proposes a two-component utility-based approach to optimize the dose selection process in order to maximize the expected utility in phase III. The first component is for efficacy and the second component is for safety. The choice of the utility function approach was driven by Decision Theory[8] that claims that utility functions are the most natural and consistent way to describe and rank preferences or decisions.

More precisely, we consider a dose-ranging trial (phase IIb) comparing J doses of a new product versus placebo followed by a pivotal phase III trial with a single dose selected versus placebo. Efficacy is characterized by a unique continuous

endpoint which is supposed to be the same in phase II and in phase III. Safety is modelled using a binary endpoint, with "0" denoting no toxicity and "1" denoting the presence of toxicity.

This paper aims to describe and implement a novel utility-based Bayesian framework in order to select the optimal dose. The impact of the the phase IIb study sample size on the quality of dose selection and the chances of successful development is assessed, the phase III sample size being fixed.

We also evaluate the interest of performing an interim analysis when half of the patients are enrolled (sequential designs). The purpose is then to assess whether or not, for large phase II trials, allowing the possibility of choosing the dose in the middle of the study and continuing the study to the end if the interim analysis is not conclusive, could reduce the size of the phase II trial while preserving the relevance of the final dose choice[9].

The paper is organized as follows. In Section 2, mathematical formalization of the dose-response modelling approach will be presented, along with the decision-making framework including efficacy Probability of Success (PoS) computations (Success being defined as significant comparison of the selected dose versus placebo in the phase III trial), the sponsor's strategy to choose the optimal dose, the Go/NoGo decision rules and the decision criteria/rules to stop at the interim analysis for futility. At the end of Section 2, we display our simulation protocol, followed by the proposed dose-response simulation scenarios. Section 3 is dedicated to results assessment and interpretation. Finally, Section 4 summarizes our Bayesian decision-making framework, addressing the proposed method, and discussing the choice of the utility function, thresholds related to decision rules, and decision criteria for interim analysis. Some perspectives are highlighted at the end of Section 4, suggesting prior assessment to guide the sponsor and improve the decisions, and advocating a re-evaluation of the choice of dose-response models, in terms of robustness, with the possibility to perform a model averaging approach.

2 Materials and Methods

Utility functions are generally introduced and defined within the decision theory framework. Utility describes the preferences of the decision maker and classifies/orders decisions. A decision theory result suggests that, in a risky environment, all decision rules can be compared and classified using the expectation of a certain function, called the utility function. This utility approach is flexible because it enables to account for: safety issues, economical/financial aspects, etc. In the literature, utility functions defined by economical/financial quantities have been considered by several authors. Sometimes, under the names of "Expected Net Present Value (ENPV))", these criteria have been used to evaluate and classify various dose selection methods, see[10,11] for example; but, in these works, they have not been used to select the doses themselves. Similar utility functions have been defined and used for design optimization of an early phase clinical trial, see[12], or to optimize (in terms of sample size and Go/NoGo decisions) a phase II/III program in an oncology context and time-to-event (death) as efficacy criterion, see[13]. Utility functions have also been used to monitor adaptive designs. In[14], the authors considered a Bayesian framework and defined trial stopping rules based on the posterior probability of a given arm to be the best arm, with respect to some utility function. Those stopping rules are similar to the one we propose in this work, but they are not focused on the dose selection problem in a dose finding trial.

Because we think that utility functions defined by economic or financial considerations (such as the cost of phase III, expected financial reward in case of successful launch of the drug) are difficult to specify with enough confidence or precision at the beginning of the drug clinical development, we preferred to focus on utility functions only defined by efficacy and safety considerations. Therefore, we propose a decision-making framework based on a new type of utility function that, following a phase II study, can drive sponsor's decision with respect to the continuation, or not, of the drug development as well as the selection of the best dose for phase III: a utility function considering simultaneously and explicitly the efficacy and safety drug profiles.

2.1 Dose-response modeling

In this paper, we chose to model efficacy via an Emax model, and safety via a Probit model. Note that in dose-finding framework, the most frequently used model for efficacy is the Emax one[15,16,7]. This model assumes a monotonic (either increasing or decreasing) dose-response. It is parameterized via the placebo effect, the maximum asymptotic effect over placebo and the ED50 dose (i.e. the dose associated to half of the maximum effect). On the other hand, we use a Probit model for safety, it directly follows from (multivariate) dichotomization of normally distributed data.

Here is the mathematical formalization of our modelling approach with all the necessary notations and calculations; assuming J dose values d to consider:

(i) The dose values d are denoted by d_j, $j = 1, ..., J$.

(ii) $Y_{d,i}$ represents the random efficacy response of patient i in dose d arm, with $i = 1, ..., n_d$, where n_d is the number of patients for the dose d in phase II study. It is assumed that $Y_{d,i} \overset{iid}{\sim} N(m(d;\theta), \sigma^2)$ where $m(d;\theta)$ is the expected mean effect of dose d, and σ is the residual variability (standard deviation of residual error). The empirical mean responses in dose d and placebo are denoted by $\bar{Y}_d$ and $\bar{Y}_0$ respectively.

(iii) N_2 and N_3 denote the phase II and planned phase III sample sizes respectively. N_3 is assumed to be constant.

(iv) For safety, we used the following Probit model: $\pi(d, \lambda) = \mathbb{P}(W = 1|d, \lambda) = \Phi(a + b \times d)$, $\lambda = (a, b)^t$, where a is the intercept parameter, b is the dose effect, W is the binary toxicity outcome for one patient, 1 for toxicity and 0 if no toxicity, and Φ is the Cumulative Distribution Function (CDF) of the standard normal distribution.

(v) For efficacy, we used the following Emax model: $m(d;\theta) = \theta_1 + \dfrac{\theta_2 \times d}{\theta_3 + d}$, $\theta = (\theta_1, \theta_2, \theta_3)^t$ is the parameter vector, where $\theta_1 = E_0$ is the placebo effect, $\theta_2 = E_{max}$ is the maximum effect compared with placebo and $\theta_3 = ED_{50}$ is the dose with half of the maximum effect.

(vi) Let $\Delta(d)$ and $\bar{\Delta}(d)$ be the expected mean difference versus placebo and its estimate, respectively. We then have $\Delta(d) = m(d;\theta) - m(0;\theta)$, $\bar{\Delta}(d) = \bar{Y}_d - \bar{Y}_0$ and $E(\bar{\Delta}(d)) = \Delta(d)$.

2.2 Decision-making framework

In the following, we will discuss our proposed utility function, as well as computations of its efficacy and toxicity components.

2.2.1 Utility function

Several utility function types can be proposed and explored through simulation scenarios (see[17]). For the specific analysis of phase II, we thought it would be interesting to consider utility functions of the following form $U = (efficacy)^h \times (safety)^k$. Efficacy is an increasing component of the dose, this term depends on the efficacy of doses, particularly on effect sizes. We chose to characterize the efficacy component as a function of the PoS, the power of a phase III trial with a fixed sample size of N_3 patients: it has the advantage of normalizing this component, ranging between 0 and 1 whatever the efficacy criterion (quantitative, binary, time to event, etc.).

The PoS can be computed using standard calculations. For the case of a balanced phase II trial, efficacy is tested as: $Z = \dfrac{\bar{\Delta}(d)}{\sqrt{2SE^2}}$, with $SE^2 = \sigma^2/(N_3/2) = 2\sigma^2/N_3$ at level α, and the power results in:

$$PoS(d, \theta) = \mathbb{P}_{H_1}(Z \geq z_{1-\alpha}) = 1 - \Phi\left(z_{1-\alpha} - \frac{\Delta(d)}{\sqrt{2SE^2}}\right).$$

Contrarily, the safety component is a decreasing term depending on the dose. This term depends on the toxicity of the doses. We chose to express it according to the probability of observing a toxicity rate (i.e. the percentage of patients having an adverse event) lower than or equal to a threshold t in the dose arm, during a phase III trial of N_3 patients in total: this also has the advantage of "normalizing" this component by varying it between 0 and 1. Note that we are evaluating here the toxicity of the dose (commonly used approach in oncology), but the method can be easily adapted to a pairwise comparison on safety between placebo and the selected dose (with some 't' non-inferiority margin for instance). The number of patients having a toxicity is a binomial distribution of parameters $N_3/2$ and $\pi(d, \lambda)$, where $\pi(d, \lambda)$ represents the probability of toxicity corresponding to dose d as defined in (iv), Section 2.1.

We considered then utility functions of the form: $U(d, \theta, \lambda) = PoS(d, \theta)^h \times \mathbb{P}(tox_{obs}(d, \lambda) \leq t)^k$, where h and k are parameters reflecting the respective contributions of efficacy and safety to the utility function: the higher the k (resp. h), the higher the penalty for safety (resp. efficacy). An alternative definition would be to set a fixed parameter, say s, such that $s = h + k$, or, equivalently, define an additional parameter w with the following two components: $h = w \times s$, $k = (1 - w) \times s$; in this latter case, both parts of the utility would have some common grounds and parameter 'w' could be used as a weight on the less relevant endpoint. Parameter t is a safety parameter controlling over toxicity in phase III.

A specific characteristic of the proposed utility function (as compared to the ones proposed in[10] or[14] for instance) is that both its efficacy and safety components depend on the sample size of the phase III study. This choice is intended to reflect real life conditions where Go/NoGo decisions and dose selection at the end of phase II always relate to the sample size the

sponsor can afford for a superiority phase III trial. This can be viewed as a pragmatic choice. Moreover, unlike the additive form of the utility function proposed in[14], where a utility function is defined as a linear combination of Response and Quit rates, the utility form in this paper considers both efficacy and safety components in a multiplicative way. This form reduces the risk of compensation between a very bad safety profile by a very good efficacy one; with such multiplicative utility form, a good profile is required on both aspects, and efficacy/safety balance is better ensured.

For the sake of simplicity, and in order to facilitate the reading, we will drop in the following of this paper the parameters in the notations of the quantities of interest when there is no ambiguity. For instance, we will note $PoS(d)$ instead of $PoS(d,\theta)$, $U(d)$ instead of $U(d,\theta,\lambda)$, etc.

2.2.2 Optimal dose and decision rules

We propose a decisional framework based on a Bayesian approach: the sponsor defines priors for the parameter estimates of both efficacy and toxicity dose-response models and define decision rules based on posterior (following analysis of the dose-finding study data) distributions of quantities of interest.

We use a MCMC approach[18,19], particularly a Metropolis-Hastings algorithm to capture the posterior of the model parameters and key quantities of interest: utility, PoS, etc. For instance, samples from the posterior of the PoS can be obtained from MCMC iterations:
$\widehat{POS}_i(d) = \Phi\left(\left(m(d;\theta^{(i)}) - m(0;\theta^{(i)}) - 1.96 \times \sqrt{2SE^2}\right) / \sqrt{2SE^2}\right)$, where $\theta^{(i)}$ is the vector of efficacy model parameters θ simulated at iteration i. The advantage of Bayesian framework over a purely frequentist approach lies in its ability to account for the uncertainty in parameter values in the decisional process and also, in allowing greater flexibility in the definition of the decision rules.

Likewise, a posterior distribution of toxicity model parameters is obtained using a MCMC approach, where $\lambda^{(i)}$ is the simulated value of the toxicity model parameter vector λ obtained at iteration i.

For each study, the sponsor makes two decisions:

(i) Identification of the recommended dose: at each MCMC iteration, one identifies the best dose as the dose with the highest utility score: for all doses d_j, we compute an MCMC estimation of $\mathbb{P}_{post}(d_j \textit{ has the highest utility})$, denoted as $\widehat{\mathbb{P}}_{post}(d_j = \textit{optimal dose}|\textit{data})$. The recommended dose $d*$ for phase III is the dose for which this probability is the highest one, i.e. the dose being the most often identified as the best one among all MCMC iterations. The details of computational aspects related to the choice of the optimal dose are given in the Supporting Information section 2. In (the unlikely) case two doses have exactly the same probability of being the best dose, the lower dose is chosen and recommended for phase III. We also compared alternative decision rules for dose selection, but the one suggested in this paper appeared to provide the best performance in terms of decision quality, see[20] for further details.

(ii) Go / NoGo decision: the sponsor computes the posterior expected PoS, denoted by $\tau = mean_{MCMC}(\widehat{PoS}(d*))$, and the posterior expected toxicity probabilities of the recommended dose $d*$, denoted by $v = mean_{MCMC}(\widehat{\mathbb{P}}(tox_{obs}(d*) \leq t))$, separately. The 'Go' for phase III is then decided if τ and v pass prefixed efficacy and toxicity thresholds denoted by threshold.eff and threshold.safe respectively. In other words, the sponsor chooses 'Go' if $\tau >$ threshold.eff and $v >$ threshold.safe. These thresholds are at the study level, they depend on the therapeutic area and the objectives of the study.

2.2.3 Sequential design

We consider also the case of a sequential design and propose an adapted utility-based decisional framework. The sequential design consists in performing an interim analysis when a fraction (for instance half, as in the simulations we preformed) of the total sample size has been enrolled: following the interim analysis, the sponsor might decide to terminate the study or to continue until the total planned sample size is enrolled. Regarding the interim analysis, we propose a simple and intuitive method: one stops at the interim analysis for efficacy if $\widehat{\mathbb{P}}[U(d*) > U(d_j) \textit{ for all the other doses } d_j|\textit{data}] \geq l$, where $l \in [0,1]$. This threshold should be high enough to guarantee accuracy of the dose choice, but not too high, otherwise frequency of early termination will be decreased and studies will be rarely terminated at interim. Details related to the computational aspects of this interim analysis criterion are given in the Supporting Information section 2. Note that an early termination of the trial at the interim analysis is not necessarily a positive outcome: we can also stop the analysis for futility, i.e. we stop at interim and we do not Go to phase III, with the same decision criteria as the ones for the fixed design (if $\tau <$ threshold.eff or $v <$ threshold.safe at interim).

2.2.4 Optimal dose estimation method: Batching approach

When implementing a MCMC approach, a subsampling method (also known as thinning) is usually adopted to remove autocorrelations, or a batching method[21,22,23,19] to avoid information loss and ensure convergence, and to possibly estimate the variance of the MCMC estimator (the latter issue was not particularly the purpose of implementing this method in this paper). Here, a batching method is implemented for a different purpose: govern the dose selection process in refining the dose selection rule mentioned in the previous section. Indeed, the main idea is to select the dose $d* = d_j$ such that $\widehat{\mathbb{P}}_{post}(d_j = optimal\ dose|data)$ has the highest value amongst all doses. Instead of simply computing standard MCMC estimates of those posterior probabilities (i.e. $\widehat{\mathbb{P}}_{post}(d_j = optimal\ dose|data)$ for all doses d_j), we will apply a batching method fully described in the Supporting Information section 2: it consists in computing, first, partial sums of the MCMC utilities iterates over batches of a sufficiently large length. With the latter method, smoother and more concentrated posterior distributions are obtained and therefore, two aspects are ensured: reducing variability and avoiding information loss within the chain. For completeness, this approach is compared, through simulations, to a traditional subsampling approach in Supporting Information section 7.4.

2.3 Simulations

In the following, we describe our simulation protocol and our chosen efficacy/safety dose-response scenarios. Note that all the chosen values stated thereafter are applied for the analysis of each simulated phase II trial and are the same in each scenario.

2.3.1 Simulation protocol

We simulated 1000 phase II studies in total. Robustness of the results was checked, by simulating 5000 phase II studies for some scenarios: simulation results were similar to those obtained with only 1000 phase II studies, which implies that the latter number of simulated studies is sufficient to guarantee precise/robust results. For each simulated trial, we made the following assumptions.

Indeed, our models can be applied to different numbers of doses (or even different dosages) but for our simulations, we consider four active doses with the following values, $d = 2, 4, 6, 8$, and one placebo with the following value, $d = 0$.

We consider informative priors for E_0 and ED_{50}, and non-informative prior for E_{max}: $E_{max} \sim N(0, 100)$, $ED_{50} \sim U[1, 10]$ and $E_0 \sim N(0, 1)$. Regarding ED_{50}, we considered this prior as it is consistent with the fact that at this stage of drug development, phase II or phase IIb, the sponsor has quantitative information (based on pre-clinical or phase I/pharmacodynamic studies) about the relevant dose range and that this reflects in design doses. Regarding the prior for E_0, we assume that, similarly, the sponsor has some information on the range of placebo effect.

The following informative prior distributions for the parameters of the Probit model are considered in this paper: intercept $a \sim N(q_{0.05}, 0.10^2)$, where $q_{0.05} \simeq -1.65$ is the normal distribution quantile which corresponds to 5% of adverse event in placebo arm, and dose effect $b \sim U[0, 1]$. The sponsor is considered here to have information on the percentage of toxicity in the placebo group (from epidemiological data, for instance), so the Probit model parameter a is centred around its true value and with limited variability: a coefficient of variation (i.e. ratio between standard deviation (0.10) and mean (-1.65) roughly equal to 6%). In real life, these choices are never completely non-informative, we often have an idea on the incidence of adverse event in the placebo arm. Concerning the slope, b, the choice of the prior was motivated by a conservative approach, assuming that the incidence of toxicity was necessarily increasing with the dose.

Sensitivity analyses were conducted in order to examine the performance of the designs with respect to different priors (by considering non informative priors for all model parameters for instance). Results were consistent with the ones obtained with the chosen priors in this paper, but needed more patients to reach similar properties and decision rule qualities (see Supporting Information section 7.1 for details). Some additional guidelines for prior elicitations are given in Section 4. Density plots of our prior dose-response distributions are also given in Supporting Information section 1.

We consider $N_3 = 1000$. In practice, phase III sample size is usually set to achieve a statistical power between 80% and 95%. It should be defined based on our understanding of the endpoint, relevant effect and what the drug might achieve. In case overwhelming efficacy is expected by the project team, a smaller phase III sample size can be envisaged as well.

Efficacy and toxicity are modelled and simulated as independent random variables to limit autocorrelation problem. Let n_{iter} be the total number of MCMC iterations: we simulate $n_{iter} = 150000$ safety and $n_{iter} = 150000$ efficacy parameters

separately, and then we combine both datasets in order to build the utility score for each dose / iteration. Among these iterations, we discarded an initial portion of the Markov chain sample so that the effect of initial values on the posterior inference is minimized: burn- in=150000/2=75000 first iterations.

PoS, toxicity component and utility are computed at each MCMC iteration level. Once utilities are estimated based on each $\theta^{(i)}$ and $\lambda^{(i)}$ (75000 estimated utilities after burn-in process, see Supporting Information section 3 for further details), we implement the batching method to compute posterior probabilities based on the estimated utilities. We consider a batch length, $B = 150$ (the choice of this value is also discussed in the Supporting Information section 3). In the final output, we will then have: $n = 75000/150 = 500$ batches, each batch representing the posterior partial mean of the utility for each dose. Sponsor will use these partial means to rank doses according to utility scores and choose the optimal one as explained in Section 2.2.2. Tables summarizing simulation results of the 1000 simulated studies are presented in Section 3, each result is an average value calculated over all phase II studies.

Regarding the Go/NoGo decision (for the fixed design), we have proposed the decision criteria based on threshold values for the PoS and for the probability of observing a toxicity rate lower than or equal to t in phase III. These values will depend on the therapeutic area and the objectives of the study; for efficacy, it could be equal to 0.30 in oncology for instance, see[24]; we tested threshold.eff = 30%, and it turned out to be too weak and not strict enough (see simulation results in Supporting Information section 5), we also tested threshold.eff = 90% which, as expected, was too restrictive and with this threshold we do not go often enough to phase III (see simulation results in Supporting Information section 6); we finally kept an intermediate threshold (moderate and reasonable) between the two (threshold.eff = 60%). So in simulations, we finally retained an efficacy decision criterion for the PoS, with 60% set as lower bound, and a safety decision criterion for the probability of observing a toxicity rate lower than or equal to t, with 50% set as lower bound. For simplicity purposes, the same threshold values are retained for the interim analysis.

We choose $l = 0.80$ for the interim analysis criterion, that is we stop at interim if:
$\widehat{\mathbb{P}}[U(d*) > U(d_j) \; for \; all \; the \; other \; doses \; d_j | data] \geq 0.80$.

The choice of this threshold is discussed in Section 7.4 ($l = 0.90$ is tested in Supporting Information section 6). We compare fixed designs ($N_2 = 250$, $N_2 = 500$ and $N_2 = 1000$ patients) with sequential designs with an interim analysis when half of the patients are enrolled ($N_2 = 500$ and $N_2 = 1000$ patients with an interim analysis at $N_2' = 250$ and $N_2' = 500$ patients respectively). Note that we also examined the performance of the designs with smaller sample sizes such as 100 patients for instance, but results related to those designs are not given in this paper (see Supporting Information section 7.4).

Here, we arbitrarily chose $t = 0.15$. Note that t value usually depends on the therapeutic area. For instance, a threshold of 0.30 (or 0.40) is more common in oncology and may vary in other areas; see Supporting Information section 7.3 for sensitivity analysis related to the choice of this threshold.

In the following, we consider $h = 1$ and $k = 2$. The choice of these parameter values is discussed in Section 4. A sensitivity analysis related to these choices is conducted in Supporting Information section 7.2.

The residual variability σ is assumed to be known and set to the value of 0.5 in the simulations. This value has been chosen in order to have, for one of our most important scenarios, named "Sigmoid" (defined in the following Section 2.3.2), a standardized effect of 0.25 for the highest dose ($d = 8$) of our design.

For each simulated phase II trial, the decision framework can be described by Figure 1.

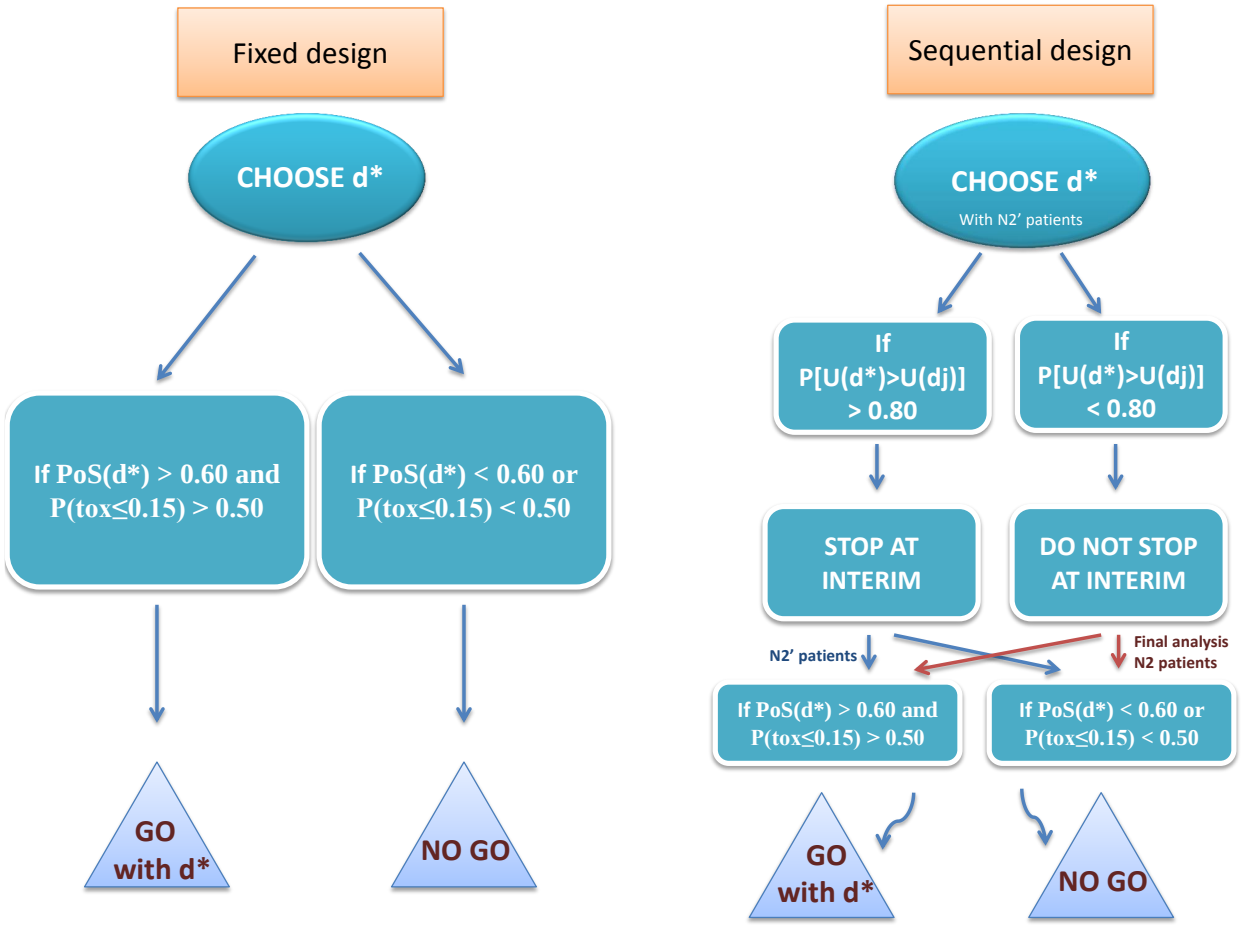

Figure 1: Schematically presentation of the decision-making framework.

2.3.2 Simulation scenarios for efficacy and toxicity

We assessed various efficacy and toxicity scenarios but for sake of simplicity, only some particular scenarios of interest are presented in this paper (see Supporting Information section 4 for additional simulation scenarios and related results).

We consider two main efficacy scenarios assumed to be the true ones reflecting the real dose-response (see Figure 2):

(i) No activity scenario: it is considered to evaluate the type I error.

(ii) Sigmoid scenario: this scenario corresponds to a smooth increase of the effect over the dose range of the design: plateau effect barely reached for the highest design dose.

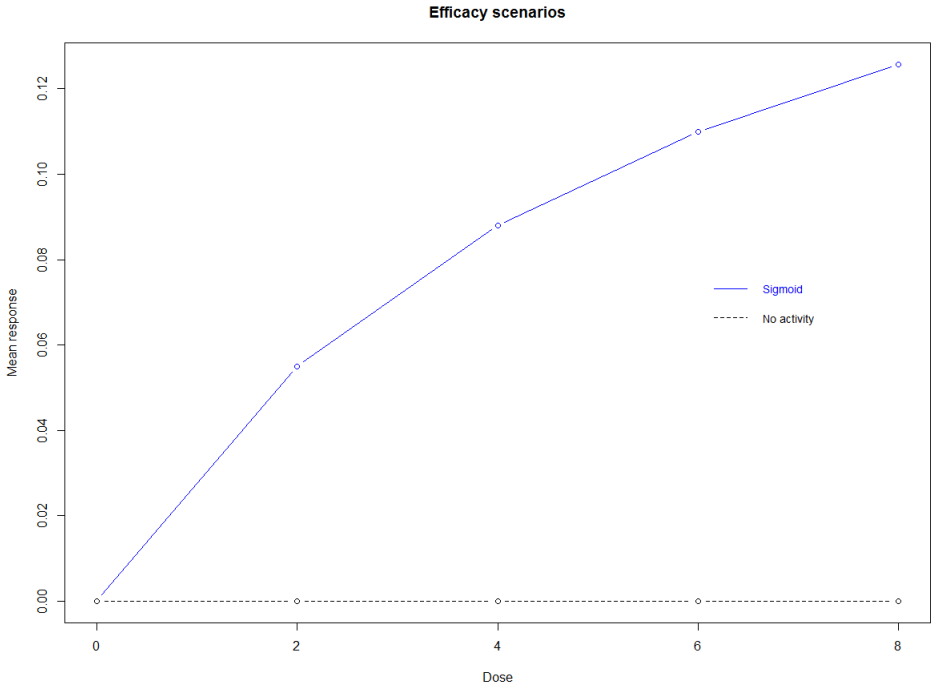

Figure 2: Overview of efficacy scenarios: mean responses as a function of d.

We also consider one main toxicity scenario (see Figure 3):
Scenario with a progressive toxicity, where the toxicity probability of the highest dose is equal to 0.20 (strictly higher than $t = 0.15$).

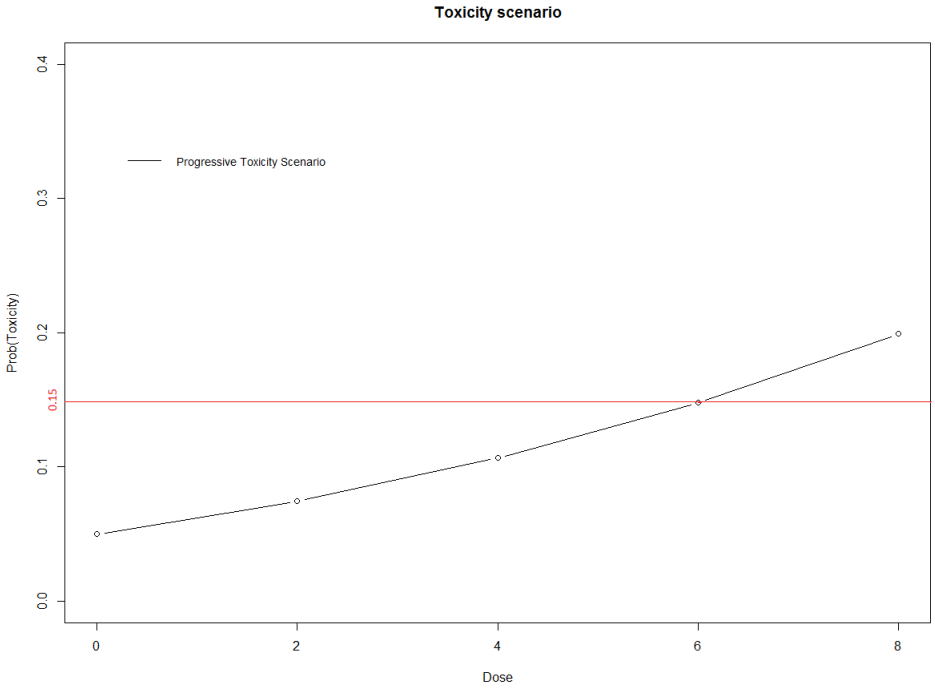

Figure 3: Overview of toxicity scenario: toxicity probability as a function of d.

3 Results

Denoting a simulation scenario by the combination of the associated efficacy and toxicity scenarios, two simulation scenarios are considered: no activity scenario $\times$ scenario with progressive toxicity (and toxicity of highest dose = 0.20), and Sigmoid scenario $\times$ scenario with progressive toxicity (and toxicity of highest dose = 0.20). For each simulation scenario, a graph highligting the corresponding theoretical curves is drawn (Figures 4 and 5), where 'Toxicity penalty' red curve represents the probability of observing more than 15% of toxicity in phase III. All the results are summarized in Tables 1 and 2. Each table contains the following:

(i) 'E(U)' is the empirical utility expectation of the chosen dose for the 1000 simulated phase II studies among 'Go' and 'NoGo' decisions (utility is set to 0 when it is a 'NoGo' decision)

(ii) 'Prob(choose(Go))' is the empirical probability of going to phase III with the chosen dose

(iii) 'Distribution selected doses (Conditional to 'Go')' represents the empirical probabilities of choosing the d=2, 4, 6 and 8 dose respectively among the 'Go'

(iv) 'Distribution selected doses (Conditional to 'Go') at interim analysis' is the empirical distribution of the chosen doses if we choose 'Go' for the interim analysis

(v) 'Distribution selected doses (Conditional to 'Go') at final analysis' is the empirical distribution of the chosen doses if we continue to the final analysis and we choose 'Go'

(vi) 'POS(conditional to 'Go')' is the empirical mean of PoSs conditional to 'Go' with the chosen dose

(vii) 'Prob(Stop at interim)' is the empirical probability of stopping at the interim analysis

(viii) '% Stop for futility' is the empirical probability of stopping for futility at interim (so this percentage is included in (vii))

(ix) 'Mean(N2)' is the mean sample size of the sequential plan

(x) 'Power' is the global power of the combined phase II / phase III program, defined as the product (ii)$\times$(vi)

3.1 No activity scenario with progressive toxicity scenario (and toxicity of highest dose = 0.20)

We started by considering a scenario with no activity to evaluate the type I error: the idea is to verify that the clinical trial stops for lack of activity, and not because of excessive toxicity. The utility function is illustrated in Figure 4 and results are given in Table 1.

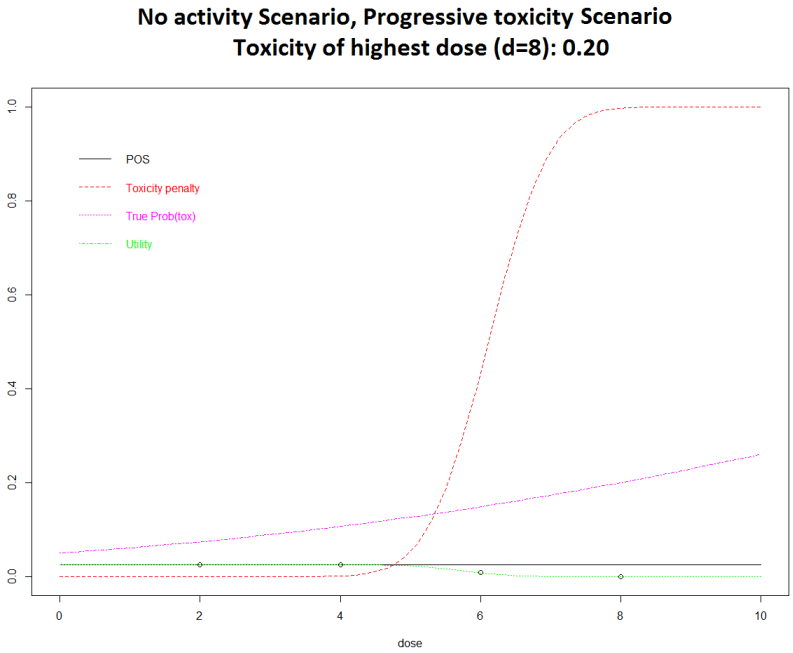

Figure 4: Theoretical curves, no activity scenario with progressive toxicity scenario (and toxicity of highest dose = 0.20).

Table 1 Simulation results, no activity scenario with progressive toxicity scenario (and toxicity of highest dose = 0.20).

No activity scenario **Progressive toxicity scenario** **Tox(d=8)=0.20** **Threshold.eff=0.60** **Threshold.safe=0.50**	N2=250	N2=500 interim at N2'=250 ; stop if : $\mathbb{P}$(best dose|data) ≥0.8	N2=500	N2=1000 interim at N2'=500 ; stop if : $\mathbb{P}$(best dose|data) ≥0.8	N2=1000
E(U)	0.003	0.002	0.001	0.001	0.000
Prob(choose(Go))	0.108	0.079	0.059	0.048	0.018
Distribution selected doses (Conditional to 'Go')	0.050 0.860 0.070 0.020	0.010 0.900 0.090 0.000	0.000 0.900 0.100 0.000	0.000 0.960 0.040 0.000	0.000 0.720 0.280 0.000
Distribution selected doses (Conditional to 'Go') at interim analysis	-	0.020 0.980 0.000 0.000	-	0 1 0 0	-
Distribution selected doses (Conditional to 'Go') at final analysis	-	0.000 0.780 0.220 0.000	-	0.000 0.670 0.330 0.000	-
POS (conditional to 'Go')	0.025	0.025	0.025	0.025	0.025
Prob(Stop at interim)	-	**0.249**	-	**0.290**	-
% Stop for futility	-	**0.202**	-	**0.248**	-
Mean(N2)=N2x(1-prob(interim))+N2'xprob(interim)	-	**438**	-	**855**	-
Power=Prob(choose(Go))xPOS(conditional to 'Go')	0.003	0.002	0.001	0.001	0.000

In this scenario, the sponsor should not decide to go to phase III since no dose is efficacious as compared to placebo. In terms of probability of wrong decision (decide to go to phase III), it is quite high ($\simeq$ 11%) with a phase II study with $N_2 = 250$ patients. But, as expected, the probability of wrong decision decreases as the sample size increases, reaching the value of approximately 2% for the largest phase II study ($N_2 = 1000$ patients). In the unfavourable case of wrong decision to go into phase III, the chosen dose is most often $d = 4$. This is due to the fact that the analysis conducted by the sponsor identifies the second dose as the highest "well tolerated" dose (based on the probability of observing more than 15% of toxicity in the phase III study). In such a scenario, the usefulness of conducting an interim analysis when half of the patients are enrolled is debatable. Indeed, the probability of stopping at interim analysis is not negligible (it is around 25% and 29% for sample size of 250 and 500 at interim, respectively) this leads to a decrease of the mean sample size of the phase II study of around 12% and 14% as compared to a fixed sample size design of 500 and 1000 patients respectively. But at the same time, even though the probability of interrupting the study and choose to go directly in phase III is small, conducting an interim analysis inflates the risk of wrongly choosing to go in phase III as compared to the fixed sample size design (risk increases from 6% to 8% with the phase II study with 500 patients and the risk increases from 2% to 5% with the phase II study with 1000 patients).

Note that if the sponsor wants to control the false go rate, it must conduct some simulations in order to adapt the decision rule so that the false go rate is maintained below an upper bound (5% or 10% for instance).

3.2 Sigmoid scenario with progressive toxicity scenario (and toxicity of highest dose = 0.20)

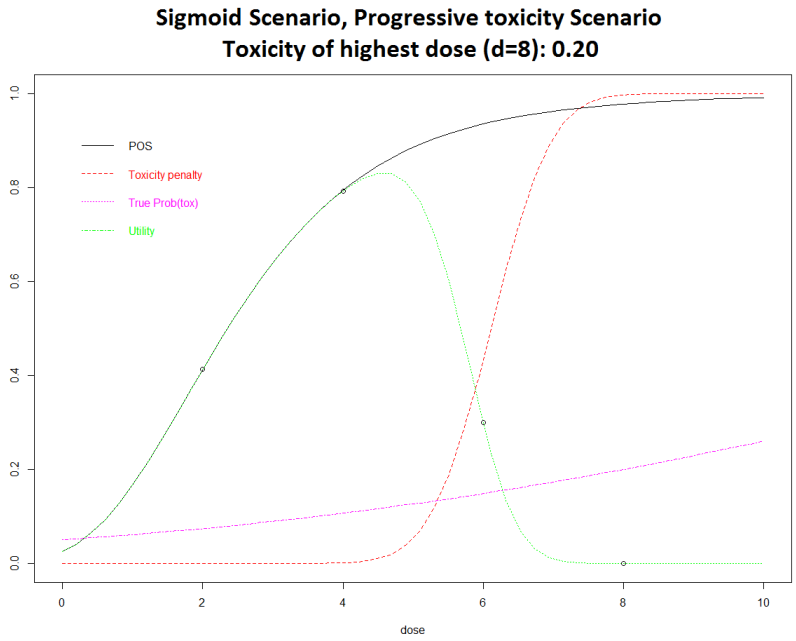

Figure 5: Theoretical curves, Sigmoid scenario with progressive toxicity scenario (and toxicity of highest dose = 0.20).

Table 2 Simulation results, Sigmoid scenario with progressive toxicity scenario (and toxicity of highest dose = 0.20).

Sigmoid **Progressive toxicity scenario** **Tox(d=8)=0.20** **Threshold.eff=0.60** **Threshold.safe=0.50**	N2=250	N2=500 interim at N2'=250 ; stop if : $\mathbb{P}$(best dose|data) ≥0.8	N2=500	N2=1000 interim at N2'=500 ; stop if : $\mathbb{P}$(best dose|data) ≥0.8	N2=1000
E(U)	0.401	0.473	0.478	0.522	0.547
Prob(choose(Go))	0.555	0.628	0.626	0.672	0.704
Distribution selected doses (Conditional to 'Go')	0.120 0.830 0.050 0.000	0.040 0.910 0.050 0.000	0.020 0.940 0.040 0.000	0.010 0.970 0.030 0.000	0.000 0.970 0.030 0.000
Distribution selected doses (Conditional to 'Go') at interim analysis	-	0.060 0.940 0.000 0.000	-	0 1 0 0	-
Distribution selected doses (Conditional to 'Go') at final analysis	-	0.030 0.890 0.080 0.000	-	0.010 0.900 0.090 0.000	-
POS (conditional to 'Go')	0.757	0.785	0.791	0.796	0.798
Prob(Stop at interim)	-	**0.404**	-	**0.655**	-
% Stop for futility	-	**0.130**	-	**0.195**	-
Mean(N2)=N2x(1-prob(interim))+N2'xprob(interim)	-	**399**	-	**673**	-
Power=Prob(choose(Go))xPOS(conditional to 'Go')	0.420	0.493	0.495	0.535	0.562

In this scenario, the optimal dose is the second one (d=4) and the true associated PoS and utility are both approximately equal to 0.8 (see Figure 5). With this scenario we can see that the probability of making the good decision (go to phase III with the second dose) is clearly dependent on the sample size of the phase II study, the probability of good decision increasing significantly with the phase II sample size. When it is equal to 250 the sponsor decides to go to phase III with a probability approximately equal to 56%, whereas the global power is equal to 42%: this sample size does not seem large enough for a sufficiently accurate estimation of both efficacy and safety models to allow good decisions. With 1000 patients, i.e. the largest phase II study, the probability of choosing to go to phase III significantly increases and reaches 70%: concomitantly, when the sample size is increased from 250 to 1000 patients, the mean utility and the global power

relative increase is approximately equal to 15% .
In terms of choice of dose, the best dose (d=4) is the selected one in most of the cases, even with only 250 patients (chosen with probability equal to 83%). But again, increasing the sample size significantly improves further the dose selection: with 500 patients in the phase II study, the best dose is selected for phase III with a probability approaching 95%.
In this scenario, performing an interim analysis when half of the patients are recruited has some interest: the probability of stopping at interim analysis is quite high (but the probability of wrong stop for futility is not negligible, equal to 13% and 20% for the interim analyses at 250 and 500 patients respectively) which leads to relative decrease of the mean sample size of 20% and 33% as compared to the fixed sample size design with 500 and 1000 patients respectively. This is interesting because this reduction of sample size does not degrade the properties of the design: considering either the probability of going to phase III, the mean utility, the selected doses or the global power, the design with an interim analysis with half patients has very similar properties as the fixed sample size design.

4 Discussion and Perspectives

4.1 Discussion

In this work, we have attached a utility value to each dose, using the utility function defined as the product of a measure of the dose efficacy and a measure of the dose toxicity. We do not claim that this utility function is necessarily the best one, but in addition to the necessary properties that should have utility functions (increase when efficacy increases while safety is fixed and decrease when toxicity increases while efficacy is fixed), it has some desirable properties: it is a smooth and a concave function (at least around the maximum of utility), this guarantees the existence of an optimal dose. Therefore, in practice, choosing a utility function $U(d)$ of the form $U(d) = (efficacy\ term(d))^h \times (safety\ term(d))^k$, with both efficacy and safety terms ranging from 0 to 1, is a pragmatic option. The choice of exponents h and k, enables to give more or less weight to the efficacy and safety terms: large values of the exponents put more constraint to the corresponding term (for instance, for a large value of h, an optimal dose should show a very high efficacy). For example, for a rare disease indication for which there is a clear unmet medical need, there should be less constraint on safety: therefore low values of k should be chosen. On the contrary, for a very competitive therapeutic area, more constraint should be put on the safety side, therefore large values of k should be chosen. In principle, a good option for the sponsor for choosing the utility function, could be to gather some experts that would rank some typical efficacy/safety profiles, those reference rankings being then used by the sponsor to choose a consistent utility function. A possible way to calibrate these values is to adopt the Delphi method[25], which is a forecasting process framework based on the results of several rounds of questionnaires sent to a panel of experts. Several rounds of questionnaires are sent out, and the anonymous responses are aggregated and shared with the group after each round.

In this work, we have considered the Bayesian framework for the statistical analysis of the phase II data. We advocate for a Bayesian approach as we think it is a more flexible framework for specifying the decision rules. We have proposed a sponsor's decision rule based on the posterior probabilities of the doses to be the optimal one: the chosen dose being the one that maximizes this posterior probability; we think that such a rule better accounts for the uncertainty in the parameter values than criteria based on the ordering of numerical "estimates" of the utilities (like the posterior mean of the utilities for instance). Efficacy and toxicity were modeled and simulated as independent random variables: in fact, in most of applications, at least apart from oncology indications, efficacy and safety variables are analysed separately, implicitly assuming weak or no correlation between the two. However, even though the utility function is defined as the product of an efficacy component and a safety component, the Bayesian analysis of the efficacy and safety variables can be jointly conducted, introducing some correlations in the posterior distributions of the efficacy and safety parameters.

Apart from the identification of the best dose, the choice to continue to phase III is a key decision. We have proposed criteria based on threshold values for the PoS, with 60% set as lower bound, and for the probability of observing a toxicity rate lower than or equal to t, with 50% set as lower bound. These thresholds have to be determined by the sponsor: for the proposed scenarios, they appeared as a good compromise between the probability of stopping in case of non interesting profile and the probability of going to phase III in case of favourable profile. In practice, to apply the methodology, the sponsor should conduct some simulations to identify the most relevant efficacy and safety thresholds for the targeted, or expected, drug profile.

In order to assess the properties of the sponsor's decision-making process mentioned above, we have conducted some simulations (1000 study replicates) under various safety and efficacy profiles and several sample sizes of the phase II study (250, 500 and 1000 patients). The quality of the decision rules were assessed in the light of the frequency, amongst the 1000 study replicates, of the good decisions either for the Go/NoGo decision or the choice of the dose for the phase III. The simulations show that estimating an optimal dose is a difficult and demanding task. For instance, for most of the

scenarios with a satisfactory efficacy profile, the probability of making the choice of going to phase III following a phase II study with 250 patients was always less than 60%, except in the scenario in which the drug shows almost no toxicity. This is due to the fact that, with this sample size, the posterior distributions of the utilities, for each of the doses, are not sufficiently concentrated around the true utility values. This leads, often, to imprecise estimations of the posterior probabilities of the dose with the highest utility score (computed for all doses d_j), which are the quantities used for dose selection, and then to wrong selection of the optimal dose. As expected, these probabilities of making the good decision increase with the sample size, but even with the largest sample size, the probability of making the good decision with a large phase II study of 1000 patients only reaches 80% when the drug does not show any toxicity: this is related to the Go/NoGo decision rule, and in this case, the sponsor should be aware of the low toxicity via simulations and should therefore adjust the efficacy and/or toxicity thresholds to increase the probability of making the good decision with fewer patients. The simulations clearly show that, regardless of time and budget constraints, the sponsor has always interest in running large phase II studies to make accurate decisions regarding the termination of the development program or the selection of the dose. But, in practice, the sample size of the phase II study is necessarily limited by budget and time constraints: those simulations show that for some efficacy and safety profiles, for phase II study of reasonable size (i.e. 250 patients), the probability of making erroneous decision (like wrongly terminate the drug development in phase II) is not negligible (varies between 35% and 44%), especially if inadequate choices of efficacy/toxicity thresholds are made, as it is the case here.

Concerning the dose selection, the probability of selecting the right dose (conditional on sponsor's decision to go to phase III) also increases as the sample size increases. For those efficacy and safety profiles that show a clear peak of utility value for one given dose, accurate dose selection can be achieved with limited sample size. In case several adjacent doses show similar utility values, the identification of the optimum dose is more challenging and requires more patients.

An important point is the assessment of the type I error, in order to verify that the clinical trial stops for lack of activity, and not because of excessive toxicity. It appears that this probability can be as high as 11% for the smallest phase II study of the fixed design (see Table 1). But again, this probability of false decision decreases as the sample size increases. Regarding the sequential designs, this probability does not exceed 8% when an interim analysis is conducted with $N_2' = 250$ patients, and only reaches 5% at most, when an interim analysis is conducted with $N_2' = 500$ patients, which globally implies a stricter control of the type I error. The efficacy and safety thresholds we have used to specify the utility functions and decision rules can be determined and calibrated by the sponsor in order to maintain the type I error below a desired level. In order to improve this type I error, the sponsor should conduct some simulations to identify the most relevant efficacy and safety thresholds for the targeted, or expected, drug profile, as previously discussed. In fact, a bad choice of these thresholds can lead to an undesired increase in the type I error.

We have seen that for some safety and efficacy profiles, it is necessary to run a large phase II study to make good decisions, whereas for others, a phase II study of moderate sample size is sufficient to make decisions with acceptable risk of mistakes (including type II error), between 25% and 35%, in other words, with acceptable phase II power, between 65% and 75% (success rate of phase II is usually between 40%-50%). An appealing strategy could be to plan upfront a large sample phase II study and perform an interim analysis, when half of the patients are enrolled, and try to make the selection at this stage. For some scenarios, in particular when the best dose shows a clear benefit in utility as compared to the others, this approach has good properties: with a quite large probability of study termination at interim analysis, it enables to reduce the sample size while maintaining the properties of the fixed large sample size design. For some other scenarios, it is less useful as the study is rarely terminated at the interim analysis, the sponsor being unable to clearly identify the best dose at interim analysis. This could be seen as a safe approach aiming to choose the optimal dose when half of the patients are enrolled, only if these analyses are reliable and clearly identify this dose as the best one among the others. In all the chosen scenarios, the sponsor decides to stop the trial when at interim analysis, $\widehat{\mathbb{P}}[U(d*) > U(d_j) \; for \; all \; the \; other \; doses \; d_j | data] \geq l$. This threshold of l has to be chosen by the sponsor: we tested several values and the threshold of $l = 0.80$ seemed to show the best compromise between quality of dose selection (with a high threshold the choice of dose is more accurate) and frequency of early termination (with a too high threshold the studies are rarely terminated at interim analysis which reduces the interest of the method). Also, in our simulations, we concluded that those interim analyses only slightly increased the risk of wrongly taking the decision to go to phase III. For the Sigmoid scenario with a progressive toxicity profile for instance, the probability of taking the wrong decision with an interim analysis at $N_2' = 250$ only increased by 0.2% compared to the fixed design with $N_2 = 500$ (see Table 2).

4.2 Perspectives

We highly recommend further development related to the way the Bayesian analyses are conducted. Risks of wrongly taking the decision to go to phase III are illustrative of the technical difficulty of simultaneously estimating two complex dose-response models with enough accuracy to properly rank doses using a utility function combining the two. In our

simulation example, the sponsor's approach is Bayesian using informative and non-informative priors for efficacy (and informative priors for toxicity) as it is usually the case in such context. This choice was driven by the will to have a "conservative" approach leading to choose priors that minimizes "subjectivity" as compared to the information included in the data. But in practice, as long as those analyses are made for internal decision-making, the sponsor could try to leverage the information available before the phase II was conducted to improve decisions. Maybe, for further development, it would be interesting to assess (through simulations) what level of information brought by the prior would be sufficient to improve the decisions; those considerations could guide the sponsor with respect to the nature of information to collect, in pre-clinical development or phase I studies, to inform those priors and then improve the utility-based decisions and dose selections. This could be done by using more informative priors related to the available information:

(i) For efficacy, based on previous studies (like a proof of concept phase IIa trial) some information could be available related to the Emax parameter for instance: a prior $N(E_{max}, \sigma^2_{E_{max}})$ not centred on 0, with a not too much inflated variability could be used

(ii) For the safety, some precise knowledge could be available such as the probability of occurrence of toxicity in the control group information that can be translated in an informative prior on the intercept of the Probit model

In our work, the interim and final analyses are conducted the same way. But in fact, according to sponsor's objectives related to the interim analysis, they could be conducted completely differently. For instance, if the aim of the interim analysis is to assess whether the drug shows some efficacy or not (with no further objective to identify the optimal dose), then a specific decision rule could be built in relation to the efficacy of the largest dose only (for example, the decision rule could be defined as a minimal PoS in phase III for the largest dose; studies would be stopped if efficacy of the largest dose is insufficient). In that case, studies would be stopped only for futility (we only stop for failure, never for success).

On the other hand, one can (numerically) optimize a phase II design, according to other decision criteria, in different contexts, based on different dose-response models: Linear, Emax (Sigmoid or not), Logistic, and even work on what happens when the sponsor computes the utility, chooses the dose with the bad dose-response model: are there more robust models than others? Is the Model Averaging approach more interesting?

An interesting perspective to work on is to transpose our proposed utility-based approach to oncology, for a phase I/phase II clinical development. However, applying a similar approach to oncology would require some significant modification of the methodology. In general, the efficacy criterion used in phase II is different from the efficacy criterion used in phase III. Very often, Best Overall Response is the phase I or phase II criterion whereas the phase III criterion is the Progression Free Survival and/or the Overall Survival. Therefore, unless basing calculations on strong assumptions, it would be difficult to assess the PoS of a dose in phase III only based on a phase I/phase II study. Phase II oncology studies with parallel group designs (including various doses or often various dose regimen) exist, but they are rare: very often the choice of dose is based on phase I dose escalation studies. Accordingly, an interesting application of our approach would be to guide the dose escalation (choice of the next dose cohort) using a utility-based approach (note that similar approaches have been considered in bivariate toxicity/efficacy Continual Reassessment Methods in[26]). A possible approach for a phase I dose escalation study would be to define a utility function having the following form $U(d) = \mathbb{P}(Response\ rate(d) \geq \pi_1)^h \mathbb{P}(Toxicity\ rate(d) \leq \pi_2)^k$. Then, after each cohort is enrolled, an optimal dose would be chosen, and would be the dose of the next cohort (other complementary safety rules could be taken into account in addition). Such a definition of utility is only applicable if we can define probability distribution for the model parameters: the Bayesian framework is the most suitable for this purpose.

Acknowledgement We would like to thank Mr. Pascal Minini for his contribution, helpful comments and suggestions that led to an improved paper.

Conflict of Interest *The authors have declared no conflict of interest.*

References

[1] LV Sacks, HH Shamsuddin, YI Yasinskaya, K Bouri, ML Lanthier, and RE Sherman. Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000-2012. *JAMA*, 311(4):378–384, 2014.

[2] EMA. ICH E4 Dose response information to support drug registration. *ICH Harmonised Tripartite Guideline*, 1994.

[3] DH Li, JB Whitmore, W Guo, and Y Ji. Toxicity and Efficacy Probability Interval Design for Phase I Adoptive Cell Therapy Dose-Finding Clinical Trials. *Clinical Cancer Research*, 23(1):13–20, 2017.

[4] F Bretz, M Branson, and J Pinheiro. Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies. *Biometrics*, 61(3):738–748, 2005.

[5] EMA. Qualification Opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of Phase II dose finding studies under model uncertainty. *Committee for Medicinal Products for Human Use (CHMP)*, 2014.

[6] J Aouni, JN Bacro, G Toulemonde, P Colin, L Darchy, and B Sebastien. Assessing Dunnett and MCP-Mod based approaches in two-stage dose-finding trials. *Biostatistics and Health Sciences*, DOI: 10.21494/ISTE.OP.2019.0397, 2019.

[7] F Miller, O Guilbaud, and H Dette. Optimal Designs for Estimating the Interesting Part of a Dose-Effect Curve. *Journal of Biopharmaceutical Statistics*, 17(6):1097–1115, 2007.

[8] LJ Savage. The Foundations of Statistics. *Dover publications, INC.*, (ISBN-13: 978-0486623498), 1954.

[9] B Bornkamp, J Pinheiro, and F Bretz. MCPMod: An R Package for the Design and Analysis of Dose-Finding Studies. *Journal of Statistical Software*, 29(7):1–23, 2009.

[10] J Temple. Adaptive designs for Dose-finding trials. *University of Bath, Department of Mathematical Sciences*, pages 1–206, 2012.

[11] Z Antonijevic, J Pinheiro, P Fardipour, and RJ Lewis. Impact of Dose Selection Strategies Used in Phase II on the Probability of Success in Phase III. *Statistics in Biopharmaceutical Research*, 2(4):469–486, 2010.

[12] LK Foo and S Duffull. Designs to balance cost and success rate for an early phase clinical study. *Journal of Biopharmaceutical Statistics*, 27(1):148–158, 2017.

[13] M Kirchner, M Kieser, H Götte, and A Schüler. Utility-based optimization of phase II/III programs. *Statistics in Medicine*, 35(2):305–316, 2016.

[14] BJ Gajewski, SM Berry, M Quintana, M Pasnoor, M Dimachkie, L Herbelin, and R Barohn. Building efficient comparative effectiveness trials through adaptive designs, utility functions, and accrual rate optimization: finding the sweet spot. *Statistics in medicine*, 34(7):1134–1149, 2015.

[15] N Thomas, K Sweeney, and V Somayaji. Meta-Analysis of Clinical Dose–Response in a Large Drug Development Portfolio. *Statistics in Biopharmaceutical Research*, 6:302–317, 2014.

[16] E Comets. Etude de la réponse aux médicaments par la modélisation des relations dose-concentration-effet. *HAL*, Médicaments. Université Paris-Diderot - Paris VII(tel-00482970):1–84, 2010.

[17] J Aouni, JN Bacro, G Toulemonde, P Colin, and L Darchy. On the use of utility functions for optimizing phase II/phase III seamless trial designs. *Therapeutic Innovation & Regulatory Science*, Under review, 2019.

[18] DV Ravenzwaaij, P Cassey, and SD Brown. A simple introduction to markov chain monte–carlo sampling. *Psychonomic Bulletin and Review*, 25(1):143–154, 2018.

[19] CJ Geyer, C Robert, G Casella, Y Fan, SA Sisson, JS Rosenthal, RM Neal, A Gelman, K Shirley, JM Flegal, GL Jones, RV Craiu, XL Meng, M Huber, JP Hobert, E Thompson, B Caffo, D Bowman, L Eberly, SS Bassett, DV Dyk, T Park, D Higdon, CS Reese, JD Moulton, JA Vrugt, C Fox, R King, M Haran, JH Park, R Peng, F Dominici, TA Louis, S Zeger, P Fearnhead, R Levy, RJ Mislevy, JT Behrens, RB Millar, F Garip, and B Western. Handbook of markov chain monte carlo. *Chapman Hall/CRC*, (ISBN: 9781420079425):3–592, 2011.

[20] J Aouni, JN Bacro, G Toulemonde, and B Sebastien. Utility-based dose-finding in practice: some empirical contributions and recommendations. *Annals of Biostatistics & Biometric Applications*, DOI: 10.33552/ABBA.2019.03.000552, 2019.

[21] C Alexopoulos and AF Seila. Implementing the batch means method in simulation experiments. *Proceedings of the 28th Conference on Winter Simulation*, pages 214–221, 1996.

[22] GS Fishman and LS Yarberry. An implementation of the batch means method. *INFORMS Journal on Computing*, 9(3):231–318, 1997.

[23] BW Schmeiser and WT Song. Batching methods in simulation output analysis: What we know and what we don't. *Proceedings of the 28th Conference on Winter Simulation*, pages 122–127, 1996.

[24] X Paoletti, M Ezzalfani, and C Le Tourneau. Statistical controversies in clinical research: requiem for the 3 + 3 design for phase I trials. *Annals of Oncology*, 26:1808–1812, 2015.

[25] AP Verhagen, HC de Vet, RA de Bie, AG Kessels, M Boers, LM Bouter, and PG Knipschild. The Delphi List: A Criteria List for Quality Assessment of Randomized Clinical Trials for Conducting Systematic Reviews Developed by Delphi Consensus. *Journal of clinical epidemiology*, 51(12):1235–1241, 1998.

[26] V Dragalin, B Bornkamp, F Bretz, F Miller, SK Padmanabhan, N Patel, I Perevozskaya, J Pinheiro, and JR Smith. A simulation study to compare new adaptive dose-ranging designs. *Statistics in Biopharmaceutical Research*, 2:487–512, 2010.