Integrating Preclinical Insights for Adaptive Dose Escalation in Phase I Oncology Trials

Helen Barnett[1], Melanie Guhl[2], Fulvio Di Stefano[3], Donia Skanji[4], Sandrine Guilleminot[5], Gaelle Saint-Hilary[2,3], Pavel Mozgunov[2,6,*] and Marie-Karelle Riviere[2,*]

[*] *These authors contributed equally*

[1] School of Mathematical Sciences, Lancaster University, UK

[2] Department of Statistical Methodology, Saryga, France

[3] Dipartimento di Scienze Matematiche (DISMA) "Giuseppe Luigi Lagrange", Politecnico di Torino, Italy

[4] Department of Clinical Statistics, Institut de Recherches Internationales Servier, France

[5] Department of Translational Statistics, Institut de Recherches Internationales Servier, France

[6] MRC Biostatistics Unit, University of Cambridge

Corresponding Author:

Marie-Karelle Riviere, Department of Statistical Methodology, Saryga, France, marie-karelle.riviere@saryga.com

Abstract

Leveraging preclinical prior information has the potential to enhance efficiency of Phase I oncology trials, if used appropriately. In this paper, a comparison of the meta-analytic predictive (MAP) prior approach and the power prior approach is undertaken in application to a setting using the Bayesian Logistic Regression Model (BLRM). A novel methodology to determine the parameters of the two approaches, based on external data, is introduced. Moreover, the escalation with overdose control criterion, commonly used in conjunction with the BLRM, is extended via an additional criterion, which allows for less conservative escalation. It is found that the inclusion of animal data is recommended to be done using the flexible MAP prior, with a justified specification of prior exchangeability. However, the benefit over not including animal data must be weighed against the potential losses.

Keywords:

Dose Finding, Early Phase, Bayesian, Pre-clinical, Prior Specification

1 Introduction

Preclinical animal data in oncology studies are mainly used for clinical consideration such as pharmacological safety and general toxicology, for the determination of the pharmacokinetic and pharmacodynamic profile of a new drug in view of the first in human study[8], and to determine the maximum safe starting dose for the first in human trial. The first in human trial then aims to evaluate the safety profile of the novel drug in humans, and find the Maximum Tolerated Dose (MTD). Within this trial, various dose levels are considered, and the design of the trial guides both the assignment of doses to patients within the trial and the final recommendation of the MTD.

There are various approaches to designing such a Phase I dose escalation study, with designs categorized into the algorithm-based, model-assisted (also known as model-free) and model-based approaches. The algorithm-based approaches whereby the design follows a pre-specified

set of rules, whilst commonly used, have been shown to be inefficient in establishing the MTD [16]. Model-assisted designs use a statistical model to guide the escalation, without making assumptions about the dose-toxicity relationship[38, 3]. Model-based designs make use of the data on each of the doses to estimate the dose-toxicity relationship using a statistical model [7]. Preclinical data provide an important basis for the design of first in human oncology trials: the starting dose is routinely justified by the observations made on animal data and this justification is often required by regulators to figure in the protocol. Yet the role of preclinical data in dose escalation remains mainly limited, and this represents a suboptimal use of the accumulated data at hand. ~~Yet, their role in dose escalation remains mainly limited to the determination of the starting dose~~. By examining how these data can be more thoroughly integrated, we aim to assess whether this approach could facilitate more efficient dose escalation strategies, while maintaining patient safety. The main challenges of this integration are to ensure a safe translation of the dose-toxicity relationship from animal to human, and the discard of preclinical data in case of a discordance with the observed clinical data. This study addresses a larger need to formally assess the performance of historical data inclusion, otherwise commonly used in confirmatory trials, when applied to early phase studies. This is a topic of broad interest, as highlighted by the recommendation of the Optimus project initiated by the FDA Oncology Center of Excellence to "develop strategies for dose finding and dose optimization that leverages nonclinical and clinical data in dose selection". The model-based designs are the most natural category to make use of the pre-clinical animal data, since these are the only type of design to make use of the actual doses.

Given the limited sample size of a typical first-in-human study, the animal data could be used more formally in the design itself via borrowing methods to help in determining the safety profile of new compounds. In this sense, several approaches have been proposed to incorporate historical data into new studies[35, 36] and some studies have already made use of historical controls in exploratory, or even confirmatory, trials[17, 18, 30]. Among them, the meta-analytic predictive (MAP) approach[24, 31, 32] and the power prior approach[5, 6, 14] are generally the most widely used. These two Bayesian approaches account for the variation across the

different sources of evidence by assigning some 'weight' to their contribution to the combined evidence. The fundamental difference between the two methods is the way in which they treat a possible inconsistency of the new study data with the historical data, so-called prior data conflict[21] or drift [36]. This inconsistency could bias the assessment of the estimated effects, leading to incorrect decisions (e.g. pursuing the development of ineffective therapies or incorrectly terminating the trial). The power prior discounts the historical data by elevating their likelihood to a certain power, whereas the robustified version of the MAP approach consists of a hierarchical model with a prior defined as a mixture between the distribution obtained from historical data and a weakly informative prior.

In both methods, the challenge is how to determine the appropriate parameters[24, 31, 32, 5, 14, 23, 12], in particular the prior confidence given in the historical data. Such choices will impact the decision making in the trial and therefore must be considered in the context of both clinical judgements regarding the relevance of the historical data to the current trial and potential operating characteristics. Since the animal data will be used to guide the dose-escalation, it must be included at an appropriate level so as to be useful but not to overpower the human trial data.

In this work, different methodologies for the incorporation of toxicological animal data in the phase I model-based dose-escalation design are explored. The power and MAP prior are compared to investigate the robustness of the informative prior they build from animal data. Both approaches involve parameters that must be pre-defined. In previous studies, these parameters were defined arbitrarily or based on expert knowledge. Here, we propose to take into account the concordance between the animal and human studies, and we propose a new way to define ~~to account for the concordance between the animal and human studies. Therefore here we present a new proposal on the definition of~~ the parameters for the two approaches by using external data[27]. A motivating case study is introduced in Section 2, to give the setting for this work. The different methodologies to be compared are presented in Section 3, then applied in simulation studies in Section 4, with an assessment of their operating characteristics. Finally, a discussion and some concluding remarks are provided in Section 5.

2 Motivating Case Study in Oncology

This work is motivated by an Oncology case study where the objective of the Phase I dose escalation is to evaluate the safety profile of a new drug and to determine its MTD, the dose associated with a target level of toxicity. Approximately 30 patients were initially considered to be enrolled in the dose escalation study, entering in cohorts of 3 - 6 patients. Each patient within a cohort receives the same dose of the novel drug, with both this dosing decision and recommendation of the MTD(s) and/or the recommended dose for expansion (RDE) determined by the statistical design, the Bayesian Logistic Regression Model (BLRM) with overdose control (EWOC).

Pre-clinical animal data is available from two Good Laboratory Practice (GLP) repeat dose toxicity studies, one study on rats and one study on monkeys illustrated in Table 1. For confidentiality reasons, the data presented here are simulated, but are indicative of the type of pre-clinical data available for such a study.

Species	**Dose** (mg/kg)	**Toxicities**	**Number of Animals**	**Study ID**
Rat	7.5	12	20	1
Rat	15	15	20	1
Rat	30	32	32	1
Monkey	3	0	6	2
Monkey	7.5	4	6	2
Monkey	15	10	10	2

Table 1: Data from pre-clinical animal studies.

Informed by the toxicology studies, in this motivating case study, a range of fixed unit doses are planned for testing, from 25 - 1400 mg with various intermediate dose levels. The starting dose is selected according to the International Conference Harmonization (ICH) S9 guidelines for choosing a starting dose for a first-in-human trial conducted in patients with cancer[8] as the 50 mg dose.

Given the small sample size available for the Phase I study and the availability of pre-clinical toxicological data from rats and monkeys, there is interest for the incorporation of the evidence

from these pre-clinical data to make dose-escalation decisions and final recommendations. This evidence should be used in such a way that the concordance between dose-toxicity in humans and animal species is considered in a pre-specified way, prior to the first-in-human trial.

3 Methods

In this study, the Bayesian Logistic Regression Model (BLRM)[22] is used for a number of reasons. Firstly, due to the suitability of model-based designs to the incorporation of pre-clinical animal data, it is deemed appropriate to use a model-based design. This model-based approach to dose-escalation models the dose-response relationship via logistic regression, and allows for flexibility in the specification of the prior distributions of the model parameters. Secondly, the two-parameter model itself is more versatile than the traditional CRM [29] since both the intercept and slope allow for the dose-toxicity model to be more flexible. It has been shown previously to merit strong operating characteristics in a number of settings [15, 25]. Thirdly, since it uses the actual dose values as opposed to a traditional skeleton, the translation of pre-clinical to clinical is both meaningful and interpretable. Finally, this is the same statistical design as used in the motivating case study, as the use of such model-based designs in practice is increasing ~~with the increasing application of such model based designs being used in practice~~ [2, 20]. Since the interest of this work is in regard to the incorporation of the pre-clinical data, for the reasons stated above, it is not deemed appropriate to compare alternative models such as the CRM or BOIN design here.

In this section, we define the methods used to incorporate the pre-clinical animal data, as well as introducing the novelty of this work: a new ~~novel~~ methodology to determine the parameters of these methods based on the concordance between human and animal studies.

3.1 Dose Escalation Procedure According to the Bayesian Logistic Regression Model (BLRM)

Here we introduce the escalation scheme and notation used for the BLRM. Following the motivating example, each of the methods proceeds along the escalation in the subsequent way. After each cohort of patients is completed, the dose recommendation for the next cohort is based on the probability that the true dose limiting toxicity (DLT) rate for each dose lies in one of the following categories: [0,16%] under-dosing, (16%,33%) targeted toxicity, [33%,100%] excessive toxicity. The choice of the thresholds has been carefully determined and is a fairly common choice in this kind of trials.

For extended safety, it is prescribed that doses for the next cohort will not be more than doubled, independently from the escalation criteria used. Dose escalation will continue until one of the two following conditions are met: the maximum sample size is reached or all doses are declared overtoxic. It can be noted that the dose recommended by the model at any stage of the trial is based on the entire history of all available DLT information from previous cohorts, as opposed to only the number of DLTs observed in the last group of patients. The MTD will be chosen at the end of the trial as the dose with the highest probability of targeted toxicity.

The first-in-human study is indexed by i^*, and investigates J_{i^*} doses. For each of the methods considered here, the following BLRM is used:

$$\begin{aligned} r_{i^*j}|p_{i^*j}, n_{i^*j} &\sim Binomial(n_{i^*j}, p_{i^*j}), \text{ for } j = 1, ...J_{i^*} \\ \text{logit}(p_{i^*j}) &= \theta_{1i^*} + exp(\theta_{2i^*})\log(d_{i^*j}/d_{Ref}), \end{aligned} \tag{1}$$

where doses are labelled d_{i^*j} for $j = 1, ...J_{i^*}$, with reference dose labelled d_{Ref} . The probability of DLT at dose d_{i^*j} is given by p_{i^*j} , with r_{i^*j} DLTs observed out of n_{i^*j} patients. The relationship between p_{i^*j} and d_{i^*j} is defined by the parameter $\boldsymbol{\theta_{i^*}} = (\theta_{1i^*}, \theta_{2i^*})$. In the following, we consider a variety of prior distributions for $\boldsymbol{\theta_{i^*}}$: fixed multivariate normal (MVN) prior, power prior and meta-analytic predictive (MAP) prior. The former does not use pre-clinical

animal data whilst the latter two do, with further details on each prior given in the subsequent sections.

Specific attention should be given to the overdose control in such trials, since the BLRM suggests the next dose to be tested as the one which has the highest probability of target toxicity, this may result in an overly aggressive escalation[22].

A common approach is to apply the dose escalation with overdose control (EWOC) [1] criterion, which restricts assignment to only doses that meet the following criterion:

$$d_{i^*j} \in \mathcal{D}_{i^*} : \mathrm{P}(p_{i^*j} > 0.33|\boldsymbol{D}) \leq \epsilon \tag{2}$$

where $\boldsymbol{D}$ is the data collected up to the current point in the phase I trial. This choice suggested by Neuenschwander et al.[22] is widely used in clinical trials which implement the BLRM, with the value for ϵ determining the strictness of the criteria.

It has been claimed by critics that this criterion, which is constructed taking into account just the overdose probability, may result in an excessively conservative escalation[39], especially in the case of unexplored doses. Therefore in addition to the EWOC criterion, in this exploration of approaches, a so called additional criterion is applied. This additional criterion is to counteract the conservative escalation of the EWOC criterion in cases where doses are unexplored.

The additional criterion states that if dose d_{i*j+1} is unexplored and dose d_{i*j} is considered safe according to the EWOC criterion, i.e. $P(p_{i^*j} > 0.33|\boldsymbol{D}) \leq \epsilon$, then the next cohort will be escalated to dose d_{i*j+1}, regardless of the EWOC criterion for dose d_{i*j+1}. Practically, after each cohort, the additional criterion will be checked and, if met, will imply escalation to dose d_{i^*j+1}. The full dose-escalation procedure is outlined below, defining the decision process after cohort k has been assigned to dose d_{i*j} and all responses from this cohort have been observed.

1. Responses from cohort k treated on dose d_{i*j} are used to update the posterior distribution of $\boldsymbol{\theta_{i*}}$.

2. Is the additional criterion met?

   Yes: Assign cohort $k+1$ to dose d_{i*j+1} (unless $j = J$, in which case assign cohort $k+1$

to dose d_{i*J})

No: Assign cohort $k+1$ to the dose in the set defined as safe by the EWOC criterion (2) that has maximum posterior probability of $p_{ij} \in (0.16, 0.33)$.

3. Repeat steps 1 and 2 for each cohort until either the maximum sample size is reached, or the lowest dose is deemed unsafe by the EWOC criterion (2).

This procedure is identical for all of the approaches considered in this work.

3.2 Pre-clinical Data

Pre-clinical data, labelled $\boldsymbol{D}_0$, consist of a certain number of studies I, which have been previously conducted on certain animal species. Guidelines (ICH M3[10], FDA M3[34], EMA repeated dose toxicity[9]) recommend the use of two animal species in toxicological studies, one rodent and one non-rodent, to ensure good coverage of potential toxic responses. As these recommendations are widely adopted worldwide, these data should systematically be available for all drug development. A single animal species has been treated with the target drug on different dose levels $\mathcal{D}_i = \{d_{i1}, ..., d_{iJ_i}; d_{i1} < ... < d_{iJ_i}\}$ in each study $i = 1, ..., I$. Moreover, each species has undergone that single clinical study. A certain number of DLTs, r_{ij}, are observed over a certain number of tested animals, n_{ij}, for each dose d_{ij} for $j = 1, ..., J_i$.

The assumption is made that the pre-clinical studies follow the same model as the first in-human study, with the number of toxicities r_{ij} assumed to follow a binomial distribution with probability of toxicity p_{ij}, dependent on each dose tested, d_{ij} , for $j = 1, ..., J_i$ in study $i = 1, ..., I$:

$$\begin{aligned} r_{ij}|p_{ij}, n_{ij} &\sim Binomial(n_{ij}, p_{ij}) \\ \text{logit}(p_{ij}) &= \theta_{1i} + exp(\theta_{2i})\log(\delta_i d_{ij}/d_{Ref}), \end{aligned} \tag{3}$$

where δ_i is the allometric scaling factor, based on the principle of proportionality between doses and dimension of the animals in terms of body weight or body surface area[33]. Here, a random

variable with log-normal distribution defined to be consistent with the FDA recommendation on allometric scaling[11] is used to account for the inherent uncertainty of these factors. The median value corresponds to the reference value in the FDA guidelines, while the 2.5th and 97.5th percentile correspond to the associated working range. In Table 2 the distributions and the FDA reference values can be found, while the methodology to obtain these distributions is treated in detail by Zheng et al. (2019) [37].

	BW (kg)			HED (mg/kg)		HED (mg/m2)	
Species	**Reference**	**Working range**	**BSA** (m^2)	λ	ν	λ	ν
Mouse	0.02	(0.011, 0.034)	0.007	-2.562	0.298	1.050	0.283
Hamster	0.08	(0.047, 0.157)	0.016	-2.002	0.302	1.609	0.287
Rat	0.15	(0.080, 0.270)	0.025	-1.820	0.323	1.792	0.309
Ferret	0.30	(0.160, 0.540)	0.043	-1.669	0.323	1.943	0.309
Guinea pig	0.40	(0.208, 0.700)	0.050	-1.532	0.315	2.079	0.301
Rabbit	1.80	(0.900, 3.000)	0.150	-1.127	0.290	2.485	0.274
Dog	10	(5, 17)	0.500	-0.616	0.301	2.996	0.286
Monkeys	3	(1.400, 4.900)	0.250	-1.127	0.273	2.485	0.256
Marmoset	0.35	(0.140, 0.720)	0.060	-1.848	0.401	1.764	0.389
Squirrel monkey	0.60	(0.290, 0.970)	0.090	-1.715	0.269	1.897	0.252
Baboon	12	(7, 23)	0.600	-0.616	0.306	2.996	0.291
Micro-pig	20	(10, 33)	0.740	-0.315	0.284	3.297	0.268
Mini-pig	40	(25, 64)	1.140	-0.054	0.258	3.558	0.240

Table 2: Log-normal prior parameters $LN(\lambda, \nu^2)$ for species-specific allometric translational factors, using body surface area (BSA) and body weight (BW) reference and working range values from the FDA guidelines[11]. HED = Human Equivalent Dose

The methodology is here applied to a repeat dose study. However, it could equally be applied to single dosing, using the appropriate dose-response models.

3.3 Multivariate Normal (MVN) Prior

The fixed MVN prior is the standard approach given by Neuenschwander [22] when using the BLRM, with:

$$\boldsymbol{\theta_{i*}} \sim MVN(\boldsymbol{m}_r, \boldsymbol{R}_r) \tag{4}$$

No animal data is used to inform this prior, however the values of $\boldsymbol{m}_r$, the prior mean vector of length 2, and $\boldsymbol{R}_r$ the 2×2 prior covariance matrix, must be specified in advance:

$$\boldsymbol{m}_r = \begin{pmatrix} m_{r1} \\ m_{r2} \end{pmatrix} \text{ and } \boldsymbol{R}_r = \begin{pmatrix} \sigma_{r1}^2 & 0 \\ 0 & \sigma_{r2}^2 \end{pmatrix} \tag{5}$$

The choice of these values are further described in Section 4.1.1.

3.4 Power Prior

The power prior uses animal data to inform the first-in human trial. The motivation of the power prior approach[5, 6, 14] is to add variability to (and therefore robustify) the animal data by elevating their likelihood $\mathcal{L}(\boldsymbol{\theta}_i|\boldsymbol{D}_{0i})$ to a given power $0 \leq \alpha_i \leq 1$ for each study $i = 1, ..., I$ in combination with the MVN distribution $\pi_r(\boldsymbol{\theta}_{i^*})$ on the parameters for the first-in-human study:

$$\begin{aligned} \pi(\boldsymbol{\theta}_{i^*}|\boldsymbol{D}_0, \boldsymbol{\alpha}) &\propto \prod_{i=1}^{I} \mathcal{L}(\boldsymbol{\theta}_i|\boldsymbol{D}_{0i})^{\alpha_i} \pi_r(\boldsymbol{\theta}_{i^*}) \\ \pi_r(\boldsymbol{\theta}_{i^*}) &\sim MVN(\boldsymbol{m}_r, \boldsymbol{R}_r) \end{aligned} \tag{6}$$

The determination of each exponent α_i in $\boldsymbol{\alpha}$ requires some special attention. Some approaches suggest to make use of the same data $\boldsymbol{D}_0$ to determine the exponent[14, 13, 26], since the use of a random exponent can lead to an intractable analytical calculation and a highly computationally intensive model[23, 12]. However, in this case, one would obtain a data-dependent prior that may lead to difficulties in interpreting the Bayesian model[4]. Therefore here we used fixed values for the α_i, that is these values are neither random nor data-dependent.

3.5 Meta-Analytic Predictive (MAP) Prior

The MAP prior also uses the animal data, but the approach allows for more flexibility. Here, the $\boldsymbol{\theta}_i$ (the parameters of the BLRM model for the species considered in study $i = 1, ..., I$) are

defined as:

$$\boldsymbol{\theta}_i|\boldsymbol{\mu}_i, \boldsymbol{\Psi} \sim MVN(\boldsymbol{\mu}_i, \boldsymbol{\Psi}) \tag{7}$$

for each animal study i. The covariance matrix $\boldsymbol{\Psi}$ representing the between-trial variability is defined as:

$$\boldsymbol{\Psi} = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}. \tag{8}$$

Following Zheng et al. [37] a MVN supra-species random-effects distribution for each animal study is specified:

$$\begin{gathered} \boldsymbol{\mu}_i|\boldsymbol{m}, \boldsymbol{\Sigma} \sim MVN(\boldsymbol{m}, \boldsymbol{\Sigma}) \\ \boldsymbol{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \kappa\sigma_1\sigma_2 \\ \kappa\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \end{gathered} \tag{9}$$

Each of the hyperparameters are given prior distributions as follows:

$$\begin{gathered} m_1 \sim N(v_1, s_1^2), m_2 \sim N(v_2, s_2^2) \\ \tau_1 \sim HN(z_1),\ \tau_2 \sim HN(z_2),\ \rho \sim U(-1, 1) \\ \sigma_1 \sim HN(c_1),\ \sigma_2 \sim HN(c_2),\ \kappa \sim U(-1, 1). \end{gathered} \tag{10}$$

where $HN(\sigma)$ represents the half-normal distribution, $N(0, \sigma^2)$, truncated to cover the positive values, while U represents the uniform distribution.

The $\boldsymbol{\theta_{i*}}$ are then defined accordingly:

$$\begin{aligned}
\boldsymbol{\theta_{i*}} &\sim MVN(\boldsymbol{\mu}_i, \boldsymbol{\Psi}) \text{with prior probability } w_i \\
\boldsymbol{\theta_{i*}} &\sim MVN(\boldsymbol{m}_r, \boldsymbol{R}_r) \text{with prior probability } w_r \\
\boldsymbol{m}_r &= \begin{pmatrix} m_{r1} \\ m_{r2} \end{pmatrix} \text{ and } \boldsymbol{R}_r = \begin{pmatrix} \sigma_{r1}^2 & 0 \\ 0 & \sigma_{r2}^2 \end{pmatrix}
\end{aligned} \tag{11}$$

where m_{r1}, m_{r2}, σ_{r1} and σ_{r2} are fixed values and $(\sum_{i=1}^{I} w_i) + w_r = 1$.

According to this model, the contribution of the pre-clinical studies to the first-in-human trial depends on the prior probability of exchangeability w_i for $i = 1, ..., I$. This is the prior belief that the behaviour of the drug in humans is comparable to the one of the species present in the studies. On the other hand, a certain prior probability w_r is placed on a fixed bivariate normal prior, which accounts for the fact that the behaviour of the drug in human may differ completely from the one in animals. Therefore the prior for the first-in-human trial is determined from the mixture of these distributions, determined by the prior weights w_i and w_r. The advantage of this method is that these weights are updated as more data becomes available, therefore adding flexibility.

3.6 Determination of the parameters of the models

One of the major challenges when using the MAP and power prior methods lies in determining appropriate weights for each, especially when animal pre-clinical data is used to inform decision and establish prior parametrization. Assigning these weights is critical, as it directly impacts how the historical animal data is incorporated into the analysis. In previous studies, these weights have been defined arbitrarily or based on expert knowledge. ~~To address this, w~~ We propose a straightforward approach based on assessing concordance and non-concordance, inspired by the survey by Olson et al.[27]. Our approach allows for a more informed and transparent weighting process that can improve the applicability of the MAP and power prior using pre-clinical data.

The study from Olson et al.[27] concludes that, considering rodent and non-rodent species, there is a 71% positive human toxicities concordance rate, with non-rodent alone being at 63% and rodent alone being at 43%. They also report the number of concordant and non-concordant studies for animal species, alongside the pre-clinical concordance by therapeutic class for all animal data. Since in our motivating case study introduced in Section 2, we have pre-clinical data from rats and monkeys, we consider studies in $\mathcal{S} = \{\text{rat}, \text{monkey}\}$. The proportion of concordant studies for rat is $p_R = \frac{86}{86+75}$, where 86 are concordant and 75 are non-concordant, and for monkey $p_M = \frac{41}{41+17}$, where 41 are concordant and 17 are non-concordant, respectively. Moreover, the general pre-clinical concordance from animal data in anti-cancer therapies is $w_A = 0.84$.

Using this independent source, the prior weight for the rat data in the MAP model can be calculated as $w_R = \frac{p_R}{p_R+p_M} w_A$ and the weight for the monkey data as $w_M = \frac{p_M}{p_R+p_M} w_A$. The weight for the robust component is therefore equal to $w_r = 1 - (\sum_{i=1}^{I} w_i) = 1 - w_A$ and is interpreted as the prior probability of non-concordance between pre-clinical animal data and human data in anti-cancer therapies.

The exponents for a power prior approach can be set to $\alpha_R = p_R w_A$ and $\alpha_M = p_M w_A$ for rat and monkey data, respectively, and can be further re-scaled to adjust prior effective sample size. Indeed, downsizing the exponent will reduce the importance of the animal data likelihood in the prior, therefore decreasing the prior effective sample size ~~to give an appropriate prior effective sample size~~.

3.7 Generalizations to multiple studies per species in the pre-clinical data

The models and methodologies presented are easily generalizable to multiple studies within a species in a pre-clinical data, whether by pooling all data from the same species and considering it as only one study, or by modeling the number of observed toxicities separately for each species

and study, if expert knowledge suggest that each study might have a different similarity with the human data (e.g. if they focus on distinct subgroups). By introducing a new index $k_i = 1, ..., s_i$ for the study, with s_i the total number of studies for species i, we define the BLRM model of the number of toxicities r_{ijk_i} for species i, dose j, in study k_i:

$$\begin{aligned} r_{ijk_i}|p_{ijk_i}, n_{ijk_i} &\sim Binomial(n_{ijk_i}, p_{ijk_i}) \\ \text{logit}(p_{ijk_i}) &= \theta_{1ik_i} + exp(\theta_{2ik_i})\log(\delta_i d_{ijk_i}/d_{Ref}), \end{aligned} \tag{12}$$

where δ_i is the allometric scaling factor for species i, common to all studies $k_i = 1, ..., s_i$.

In the power prior approach, the likelihood of all studies within a species could be grouped to be elevated to the same exponent which would represent the weight of this species in the power prior. However, the weights could also be differentiated so that their relative importance in the prior reflects their specific similarity with the human data. In that case, the product and exponents in Equation 6 would have to be indexed on the studies k_i in each species i and not the species directly:

$$\pi(\boldsymbol{\theta}_{i^*}|\boldsymbol{D}_0, \boldsymbol{\alpha}) \propto \prod_{i=1}^{I} \prod_{k_i=1}^{s_i} \mathcal{L}(\boldsymbol{\theta}_i|\boldsymbol{D}_{0i})^{\alpha_{k_i}} \pi_r(\boldsymbol{\theta}_{i^*}) \tag{13}$$

In the MAP prior approach, if each study within a species were modeled separately, the hierarchical MAP prior would need an additional layer. The parameter of the BLRM model for each study would follow a multivariate normal distribution centered on θ_i, the parameter at species level, as defined in Equation 7, with a variance parameter V_i corresponding to the inter-study variability, to be defined:

$$\boldsymbol{\theta}_{ik_i}|\boldsymbol{\theta}_i, \boldsymbol{V_i} \sim MVN(\boldsymbol{\theta}_i, \boldsymbol{V_i}) \tag{14}$$

In that case, the BLRM model parameter vector at species level θ_i and the prior components of the BLRM model parameter vector in the first-human study θ_{i^*} would still be defined as in

Equations 7 and 11 respectively.

4 Simulation Studies

Motivated by the real case-study in oncology introduced in Section 2, here we present simulation studies to compare the considered methods. The analyses have been conducted using R version 4.1.3. For confidentiality reasons, the data used in this section are all simulated data.

4.1 Setup

4.1.1 Pre-clinical Animal Data and Prior Definition

Common to all of the methods considered here is the MVN component of the prior, defined by the prior mean, $\boldsymbol{m}_r$, and prior covariance matrix, $\boldsymbol{R}_r$. We consider two sets of values for these, labelled as vague and calibrated. The MAP prior traditionally uses a vague MVN part, however here both sets of prior hyperparameter values are applied to each of the three methods in order to fully understand the behaviour of the approaches.

Vague: A weakly informative prior that allows for flat to steep curves. Hyperparameters are chosen to give a large variability on the distribution of toxicity risk at each dose, with a prior median at the reference dose of 0.2, hence labelled as vague:

$$\boldsymbol{m}_r = \left(\log\left(\frac{0.2}{1-0.2}\right), 0\right), \qquad \boldsymbol{R}_r = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$$

Calibrated: Calibrated values are obtained by conducting a grid search over a number of potential candidate values. For each combination, calibration simulations are conducted in a small number of varied scenarios using only the fixed MVN prior, and choosing the combination that gives a good level of performance overall (further details are given in the Online

Supplementary Materials):

$$\boldsymbol{m}_r = \left(\log\left(\frac{0.1}{1-0.1} \right), -0.5 \right), \qquad \boldsymbol{R}_r = \begin{pmatrix} 1 & 0 \\ 0 & 0.25 \end{pmatrix}$$

Available data from pre-clinical studies on rats and monkeys are shown in Table 1, setting $\log(\delta_R) \sim N(-1.820, 0.323^2)$ and $\log(\delta_M) \sim N(-1.127, 0.273^2)$ for allometric scaling factors for the rat and monkey data, respectively, taken from Table 2.

A selection of weights for the MAP approach are considered for comparison. Three variants are considered, with the choices motivated by various factors. The novel approach specified in Section 3.6 is labelled as MAP (specified) and is defined according to the preclinical/clinical concordance. We also consider $w_R = w_M = w_r = \frac{1}{3}$, labelled MAP (1/3), which is balanced weights for the rat, monkey and robust component. Finally, we consider $w_A = w_R = \frac{1}{2}$ with $w_R = \frac{p_R}{p_R+p_M} w_A$ and $w_M = \frac{p_M}{p_R+p_M} w_A$, labelled MAP (equal v & a), to give balanced weights between the animal and robust components. The considered weights in the power prior approach are again those outlined in Section 3.6, used as is or divided by a scaling factor to explore different prior effective sample sizes ~~and the remaining values scaled in order to give an appropriate prior effective sample size~~. Table 3 displays the prior weights used in all of the methods considered.

Prior weight	**Rat**	**Monkey**
MAP (1/3)	0.33	0.33
MAP (equal v & a)	0.21	0.29
MAP (specified)	0.36	0.48
Power	0.45	0.60
Power/10	0.04	0.06
Power/100	0.004	0.006

Table 3: Values considered for the prior weights in the six approaches that use animal data. For the MAP prior, this is w_R and w_A. For the power prior, this is α_R and α_M, the exponent to which the likelihood is elevated, divided by a factor of 1, 10 or 100 to explore different prior effective sample sizes ~~this is α_1 and α_2, with the scalar values used to give appropriate prior effective sample sizes~~.

In total therefore there are fourteen specifications of prior considered in this work. Figures 1

and 2 display these priors for the vague and calibrated MVN part respectively. Each plot displays the prior mean of toxicity at each dose, as well as the probability of being in the under, target and overdosing interval.

Prior effective sample sizes for each prior considered were computed following a method suggested by Morita et al. [19] and can be found in the Online Supplementary Material.

4.1.2 Scenarios

In order to compare the approaches, simulations are conducted over a number of scenarios, shown in Table 4. These scenarios are chosen to represent a plausible range of dose responses, with scenarios 1-7 increasing in toxicity. Scenario 1 has all doses within or under target toxicity; as the scenario number increases, the placement of the MTD within the dose range decreases until scenario 7 has all doses unsafe. Scenarios 8, 9 and 10 are chosen to reflect the first in-human trial having a similar dose-toxicity relationship to one or both of the animal studies considered, with scenario 8 similar to the rat study, scenario 9 similar to the monkey study and scenario 10 a mixture of the two. In terms of agreement with the animal data, scenarios 6, 8, 9 and 10 agree strongly with the animal data, scenarios 1 and 2 disagree strongly with the animal data, and scenarios 3, 4, 5 and 7 agree somewhat with the animal data.

Scenario	**25 mg**	**50 mg**	**100 mg**	**200 mg**	**400 mg**	**800 mg**	**1400 mg**
1	0.0001	0.01	0.02	0.03	0.07	0.12	**0.20**
2	0.01	0.03	0.05	0.10	0.14	**0.28**	0.40
3	0.03	0.05	0.10	**0.18**	**0.30**	0.46	0.60
4	0.02	0.04	0.08	**0.24**	0.35	0.40	0.45
5	0.05	0.10	**0.25**	0.40	0.55	0.70	0.85
6	0.11	**0.28**	0.37	0.44	0.61	0.73	0.80
7	0.36	0.53	0.69	0.82	0.90	0.95	0.97
8	**0.21**	0.38	0.61	0.81	0.93	0.98	0.99
9	0.09	**0.17**	0.38	0.72	0.93	0.99	0.99
10	0.15	**0.27**	0.49	0.76	0.93	0.99	0.99

Table 4: Simulation scenarios with different DLT probabilities, with MTD highlighted in **bold**.

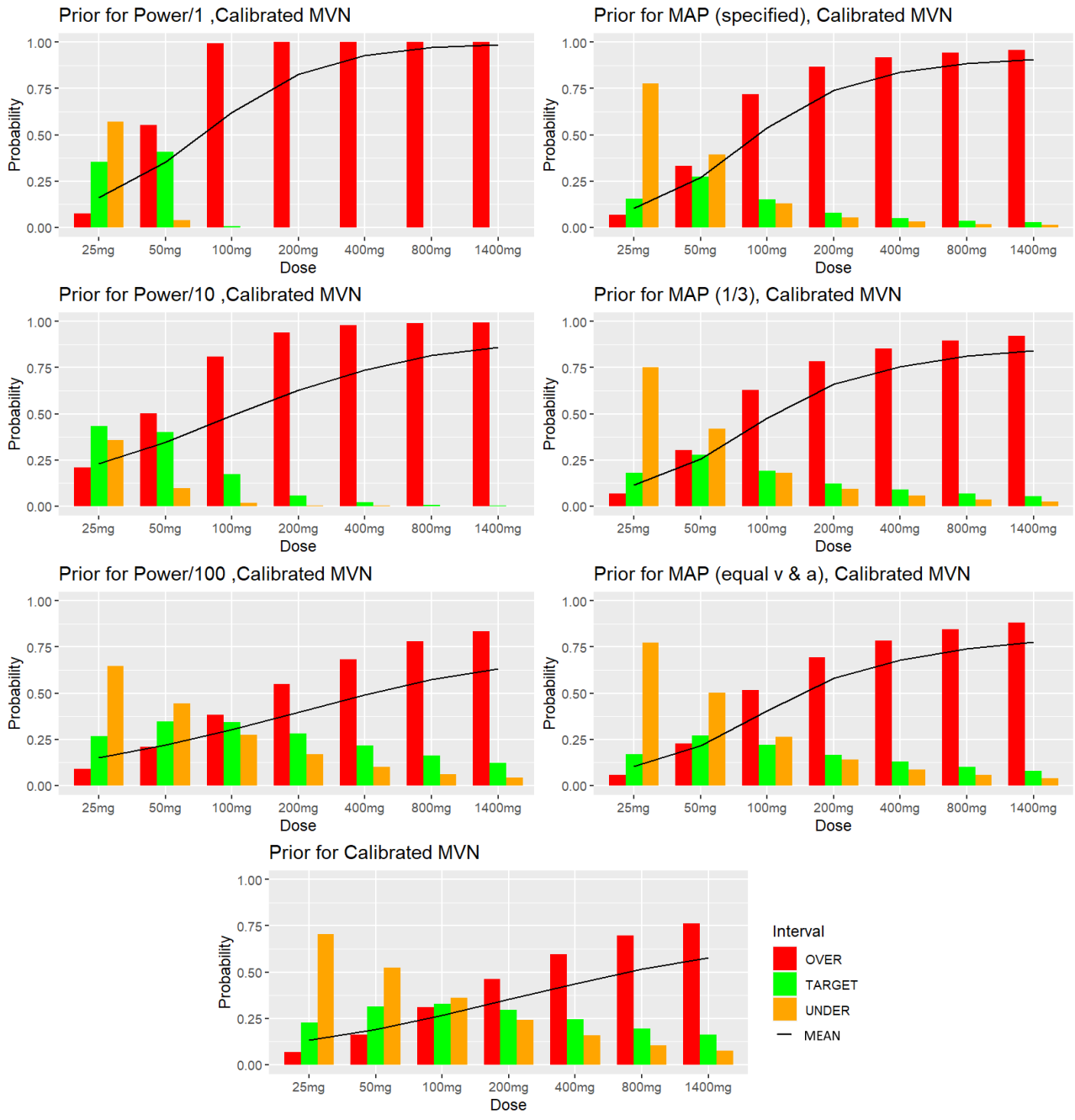

Figure 1: Prior specification for the seven priors that use the calibrated MVN part.

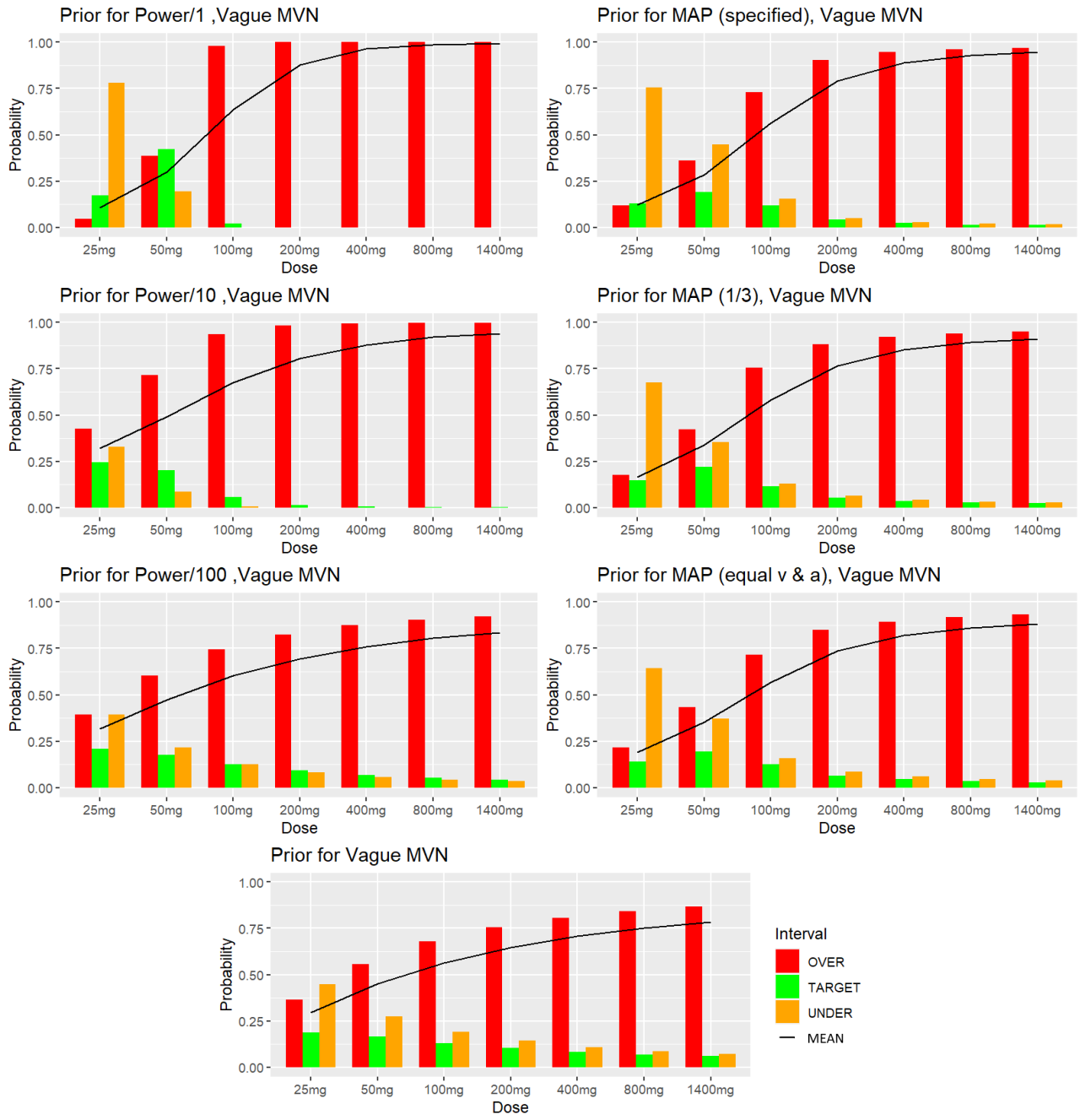

Figure 2: Prior specification for the seven priors that use the vague MVN part.

4.1.3 Operating Characteristics

The patient allocation and dose recommendation are compared across the methods, with a focus on a good performance in terms of both aspects. It is important to consider the escalation behaviour of each of the approaches in term of both the selections and assignments to ensure they are neither overly cautious or aggressive in their escalation. We report the percentage of correct selections (PCS), that is the percentage of simulations that correctly identify the MTD(s), or in the case of scenario 7, correctly determining that there is no safe dose. We also report the mean number of patients assigned to the MTD(s), to give an indication of the speed of escalation. More patients assigned to the MTD indicates that it has been identified sooner. However, this must be balanced with the recommendations and selections of overtoxic doses, where overtoxic doses are defined as those doses with a true DLT probability greater than 0.33.

4.1.4 Non-Parametric Benchmark

To provide context for the PCS, we have also included a non-parametric benchmark [28] as a comparator. This gives an idea of the difficulty of any given scenario, providing "a bound beyond which improvements are not generally possible". It does so by using the hypothetical information from all patients on all doses. As highlighted by the benchmark performance, some scenarios provide more of a challenge than others.

4.1.5 Implementation

These simulations were made in R using the Gibbs sampler of the R package *rjags*. Table 5 presents a summary of the parameters used in the simulations common to all scenarios, including the Gibbs sampler parameters and the MAP prior hyperparameters. Other parameters that are method specific (prior weights in the MAP and power prior approaches) are presented in Table 3 in Section 4.1.1. All codes used to obtain the presented results are available at `https://github.com/Saryga-SAS/preclinical_insights_dose_escalation`.

Parameter	Value	Considerations
Human weight (kg)	60	Defined by the clinical team
Set of doses (mg) Reference dose (mg) Starting dose (mg)	25, 50, 100, 200, 400, 800, 1400 25 50	Defined by the clinical team in discussion with statisticians
Escalation cap	No more than double (i.e. no more than the next dose)	Discussed by the project team given the anticipated dose-toxicity
EWOC criterion ϵ (Eq2)	0.35	Defined based on operating characteristics
Cohort size Maximum number of cohorts Maximum sample size	3 14 42	Defined based on operating characteristics
Number of simulations per scenario	1000	Chosen to have enough reassurance in the operating characteristics
Gibbs sampler: Number of chains Number of iterations per chain Burn in Thinning parameter	 4 50000 5000 10	Defined by the statistical team
MVN parameters (Eq5): m_{r1} m_{r2} σ^2_{r1} σ^2_{r2}	Vague | Calibrated $log(\frac{0.2}{1-0.2})$ | $log(\frac{0.1}{1-0.1})$ 0 | -0.5 4 | 1 1 | 0.25	Defined by the statistical team (vague: high amount of uncertainty, calibrated: high geometric mean of PCS)
MAP hyperparameters (Eq10): v_1 v_2 s_1 s_2 z_1 z_2 c_1 c_2	 m_{r1} m_{r2} 1 0.5 0.5 0.25 15 5	Defined by the statistical team depending on the desired level of informativity, following [37]
Allometric parameters: δ_R δ_M	 $\log(\delta_R) \sim N(-1.820, 0.323^2)$ $\log(\delta_M) \sim N(-1.127, 0.273^2)$	FDA guidelines

Table 5: Summary of parameters common to all methods

4.2 Results

The setup of the trial in each simulation is identical to that described in Section 2, with a maximum sample size of 42 patients and 1000 simulations performed for each scenario. Here we present figures summarizing the main results, with tables containing full information on the dose selections and dose assignments available in the Online Supplementary materials.

Figure 3 provides a comparison of the methods' PCS, with calibrated MVN in the top panel, and vague MVN in the bottom panel. The differences in the approaches are apparent across the

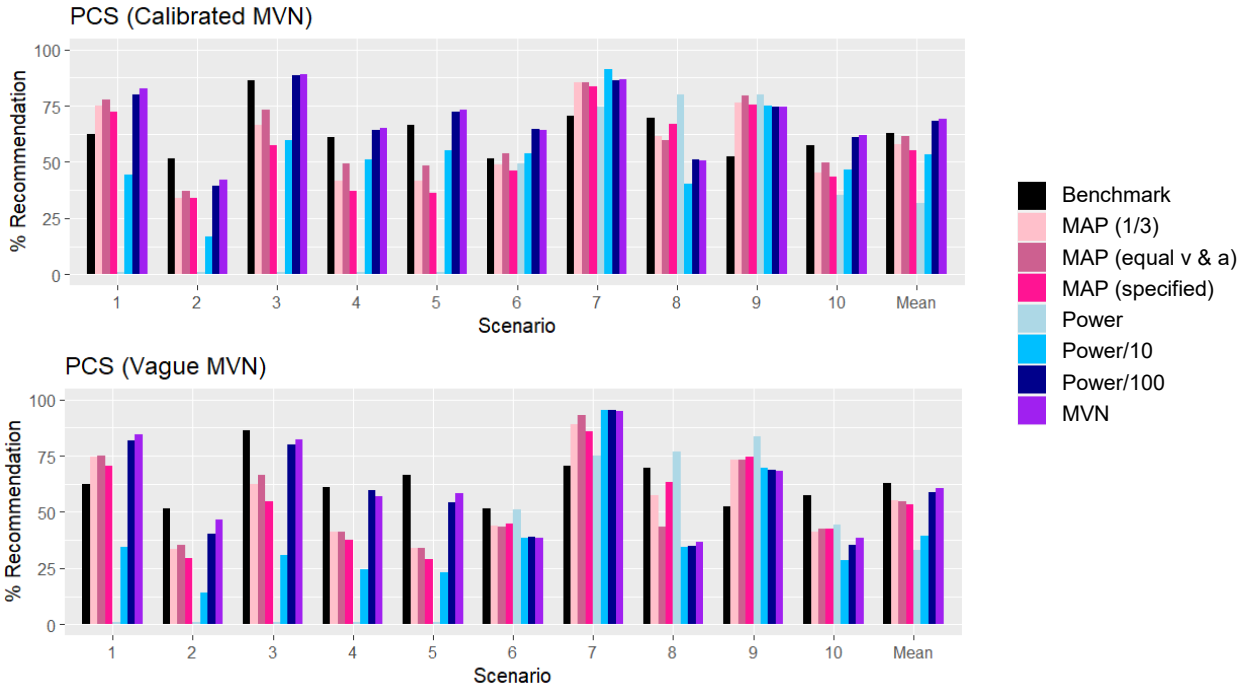

Figure 3: Percentage of Correct Selections (PCS) for all methods. The top figure uses the calibrated MVN and the bottom uses the vague MVN.

scenarios. Noticeably, the power and power/10 prior approaches perform extremely poorly in scenarios 1-5, where the human study is in disagreement with the data from the animal studies. When downweighting the power prior exponent by a factor of 100, the operating characteristics improve in terms of PCS in scenarios 1-5, and in terms of overtoxic recommendations and mean overtoxic assignments in scenario 7 where all doses are overtoxic (thus also contradicting the animal data). Given the ~~prior effective sample size and~~ inflexibility of the power prior, this is an unsurprising result, but it is important to highlight the large impact on the results. Similarly, as expected, these power priors perform well when the human study is in agreement with the data from the animal study. ~~The scaling of the powers has the expected effect that the larger the prior effective sample size, the more favour given to those scenarios where the animal and human data are in agreement, and when the animal data is down-scaled by a factor of 100, the performance is close to simply using the MVN on its own.~~ The unscaled power prior gives the larger prior effective sample size, which has the expected effect to favour the scenarios where the animal and human data are in agreement, whereas rescaling the exponent by a factor of

100 decreases the prior effective sample size and makes the performance close to simply using the MVN on its own.

The MAP prior shows more promising results due to its flexibility, with the PCS close to or exceeding the benchmark in scenarios where the animal data is both in strong agreement or strong disagreement with the human study, although still not as performant as the MVN in many cases. There is however a noticeable drop in performance in the "middle ground", where there is neither strong agreement or disagreement between animal and human data, for example in scenarios 3-6. Here, the nature of the MAP prior means that the exchangeability weights, whilst not as high as in the case of strong agreement, are high enough that the animal data still has some influence.

Despite the flexibility of the updating weights, the specification of prior exchangeability still influences the operating characteristics. On average across the scenarios, the three specifications perform very similarly, even in scenarios 6, 9 and 10, where there is a high concordance between preclinical and clinical data. However, the weights specified by the novel approach show an increased performance in scenario 8, since this is the scenario closest to the rat data, and this has the highest concordance rate. The other weights give more balanced PCS across the scenarios.

Considering the MVN prior on its own, it expectedly has the most balanced approach overall, only outperformed by the methods using animal data when there is strong agreement between the animal and human data. The calibrated version expectedly shows superior performance to the vague version, with up to 26% increase in PCS.

Figure 4 shows the percentage of overtoxic recommendations. These are small for all methods, with only the power prior ever exceeding 25% in scenario 7, where all doses are overly toxic. Of note, when the calibrated MVN is used, the MVN has more overtoxic recommendations (up to 11%), especially in scenarios where there is strong agreement between the animal and human data. This is because there is no leveraging of the animal data in the dose escalation procedure.

Figure 5 displays the mean number of overtoxic assignments, that is the mean number of patients assigned to overly toxic doses. The calibrated MVN has consistently the highest, apart

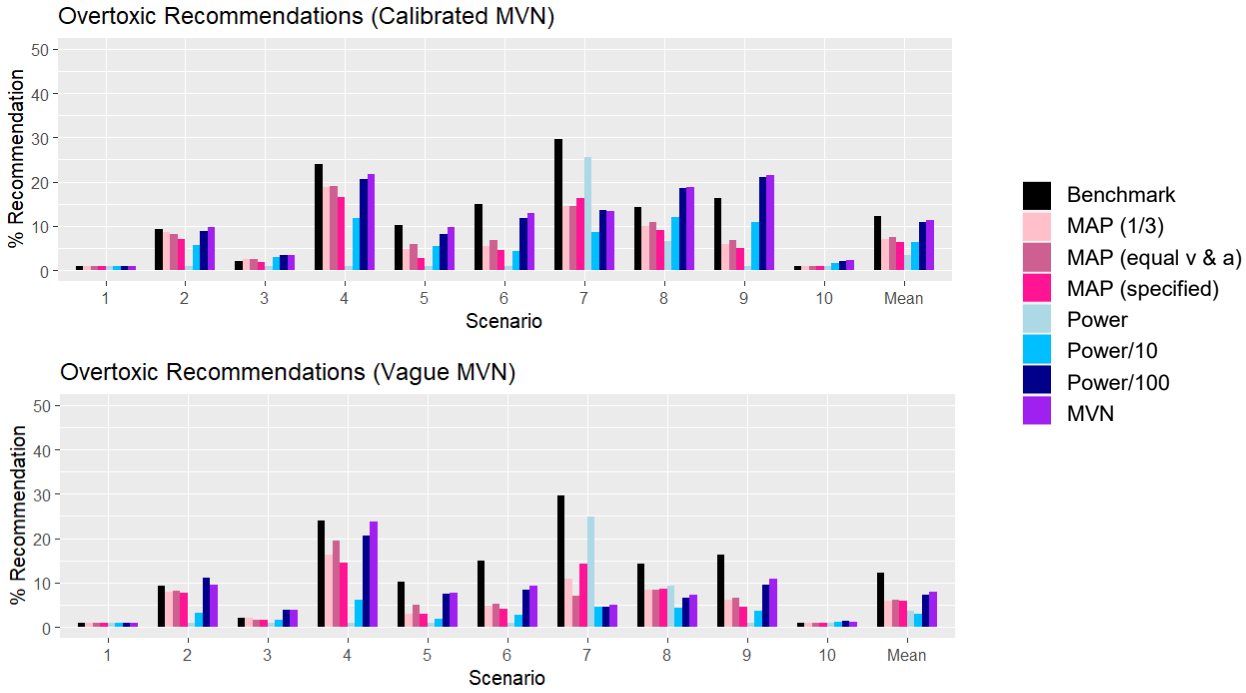

Figure 4: Percentage of Overtoxic Recommendations for all methods. The top figure uses the calibrated MVN and the bottom uses the vague MVN.

from in scenario 7, where the unscaled power prior has an average of 23 patients assigned to overtoxic doses. Here, the large prior effective sample size takes a large number of patients to overcome. Again it is noticeable that the calibrated MVN requires more exploration of the overly toxic doses to give comparable PCS to the MAP prior.

Figure 6 displays the mean number of patients assigned to the MTD in each of the scenarios. This gives an indication of speed of escalation, especially in the scenarios where the MTD is higher in the dose range. For example, it can be seen that the calibrated MVN prior on average has the quickest escalation, to be expected since it also has the highest number of overtoxic assignments. The MAP prior shows a good number of patients assigned to the MTD, more so in scenarios 8-10 where the animal data is in agreement with the toxicity scenario, but still a reasonable number in the other scenarios, especially when considering the number of overtoxic assignments. The three specifications once again perform similarly except in scenario 8, the scenario with the highest concordance rate, where MAP (specified) has more assignments to the MTD.

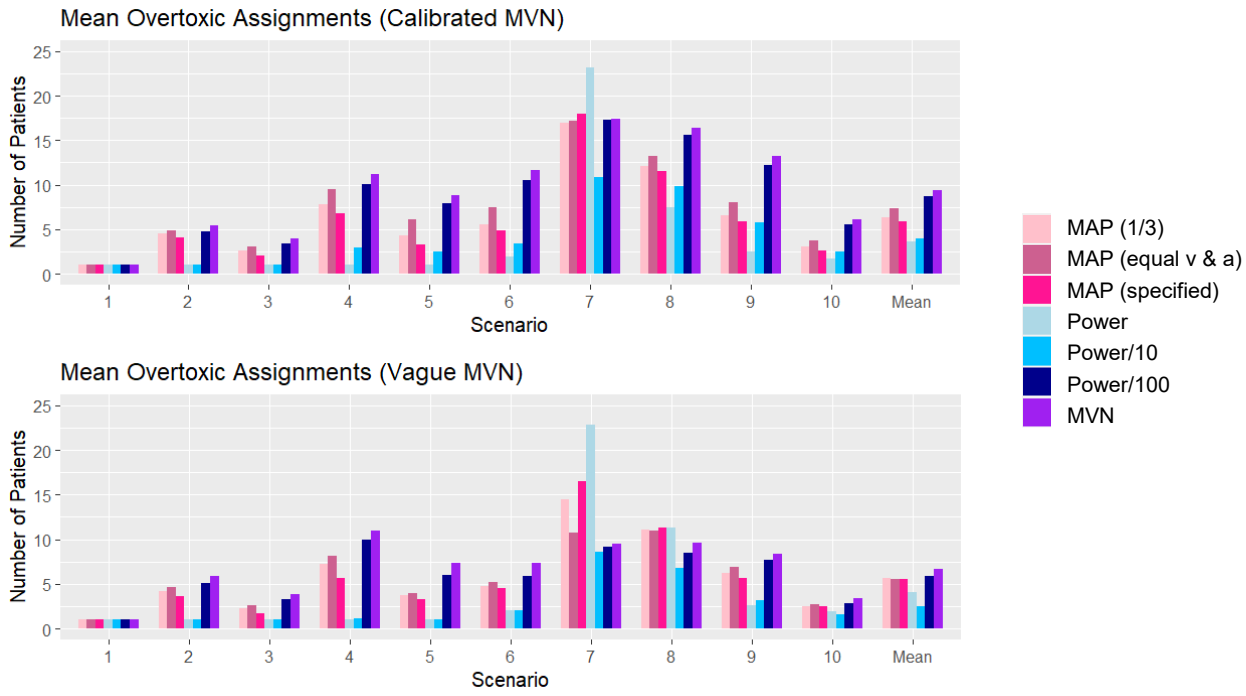

Figure 5: Mean number of Overtoxic assignments (patients) for all methods. The top figure uses the calibrated MVN and the bottom uses the vague MVN.

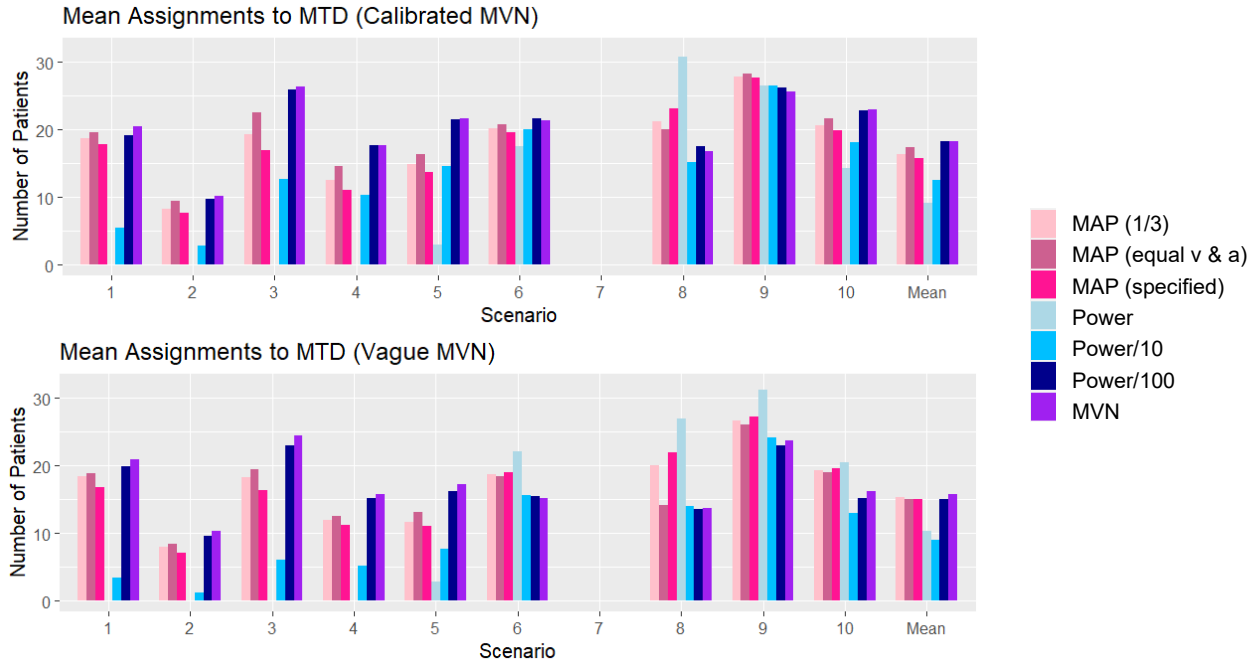

Figure 6: Mean number of assignments (patients) to the MTD for all methods. The top figure uses the calibrated MVN and the bottom uses the vague MVN.

The overall perspective of these results illustrates the benefits and downfalls of using the MAP prior, and the influence of the prior exchangeability assumptions. If the prior assumption of concordance is satisfied (i.e. scenarios 6, 8, 9 and 10 where the MTD is at 25 mg or 50 mg as in the animal data and the DLT probability distributions are similar), then the MAP prior outperforms the fixed MVN prior both in terms of PCS, with potential increases of 27%, and assignments to the MTD, with potential increases of 8 patients. However it under-performs compared to the power prior, up to 13% PCS, 4 patients more on overtoxic assignments and 7 fewer patients on the MTD. However, when this concordance is not satisfied, the MAP prior clearly holds the advantage over the power prior, with a drastic difference of PCS values of up to 78%. The MAP prior approach also holds a global advantage in terms of safety compared to the MVN prior: there are less overtoxic recommendations (down to 16%) and lower mean assignments to overtoxic doses (potential decrease of 7 patients) in all scenarios except those with the lowest MTD (scenario 7: no MTD and scenario 8: MTD at first dose). Whilst the inclusion of animal data is therefore best done using the flexible MAP prior with a justified specification of prior exchangeability since we will always be in a situation where there is uncertainty over the concordance, the benefit over not including animal data must be weighed against the potential losses in the middle ground.

When we consider mean performances across the scenarios, although the MVN offers the simplest and globally the better performing design in terms of PCS at 9% higher on average than the MAP prior, this is at the cost of on average 2 patients more on overtoxic assignments, and 3% more overtoxic recommendations. The use of animal data in the MAP prior means that less exploration of the higher doses is necessary, hence the benefit in terms of safety.

Of note, calibrating the MVN prior hyperparameters has an impact on the performance of all methods: the results of the simulation study show that using a calibrated version of the MVN prior increases the PCS and mean assignments to MTD in the MVN prior approach (up to 26% more correct selections and 7 more patients assigned to the MTD) but also in the MAP prior approach which uses the MVN distribution as its robust component (up to 24% more correction selections and 9 more patients assigned to the MTD) and the power prior

approach which multiplies its likelihood components by the MVN distribution (up to 32% more correction selections and 8 more patients assigned to the MTD). This is at the cost of more overtoxic recommendations and higher mean overtoxic assignments (which is not surprising as the calibration is done based on PCS). Indeed, in the calibrated MVN prior approach, there are up to 8 more patients assigned to overtoxic doses and up to 11% more overtoxic recommendations; in the calibrated MAP prior approach, up to 7 more patients assigned to overtoxic doses and up to 9% more overtoxic recommendations; in the calibrated power prior approach, up to 8 more patients assigned to overtoxic doses and up to 9% more overtoxic recommendations.

To investigate the robustness of these results to sample size, as dose-finding studies can imply fewer patients, these simulations were redone with a maximum sample size of 24 patients instead of 42. The results, in terms of PCS, overtoxic recommendations, mean overtoxic assignments, and mean assignments to MTD, are available in the Online Supplementary Material. As expected, this sensitivity analysis shows deteriorated results for the power prior approach when it is not downweighted, as the prior carries even more weight compared to the limited amount of data.

5 Discussion

In this paper, the incorporation of pre-clinical animal data in phase I studies is discussed, with two different methods to incorporate historical data compared. A novel approach to define the parameters of the two models, based on external concordance data, is presented.

The incorporation of historical data in clinical studies is common and well documented for confirmatory trials, but the use of animal data in early phase studies is often limited to the definition of the starting dose. The accumulated preclinical data could be more extensively used to inform the dose escalation strategy, provided that the safety of this approach was better evaluated. This work aims to encourage the incorporation of preclinical data in first in human studies by presenting two approaches, the MAP prior and power prior, and comparing

them to the classical approach, the MVN prior, that does not take historical data into account. The novel definition of the prior weights used in both historical approaches, based on observed concordance data, ensures a stronger justification of the parameter calibration than what has been done in previous work, allowing for more confidence in the approach.

The novel methodology to define the parameters for the robust MAP and power prior approach has been shown to perform well. Its main advantage relies on the potential to use external data to define the parameters for the model, which can have an easy interpretation for internal studies and regulatory approval. Though the proposed method to specify weights in the MAP prior approach does not uniformly outperform the alternatives in our simulation study, it avoids uninformed weights based on statistical consideration, instead considering available external information, which is of interest because the main challenge in the use of the MAP is the determination of appropriate parameters.

This novel methodology~~It~~ permits the comparison of the two models here presented on equal terms, which means with the same given prior probability of concordance between the animal data and the current data. Other methods used for the definition of parameters in the robust MAP approach involve the calculation of operating characteristics, which require extensive calculation in the presented case. Expert knowledge is another option, but it is highly subjective. In the case of the power prior, other methods have been proposed in the literature to select the prior exponents, but these methods require extensive numerical calculations or the use of current data in defining a parameter that should be selected "a priori". Additionally, in the proposal of the power prior method it is also noted that the use of a fixed exponent is more easily interpretable[14]. Therefore, the methodology presented in this study can be valuable in the presence of appropriate data from external sources, such as the literature or internal studies that are distinct from the ones used in the model.

The calibration of the MVN prior parameters, based on a grid search, is another novelty of our work. It allows the design of the study to be taken into account to optimize the performance in terms of PCS, instead of using the usual vague values with large variability on the distribution of toxicity risk. As seen in the results of our simulation study, this calibration allows higher PCS

and assignments to the MTD at a cost of higher overtoxic assignments and recommendations. Considering the overall small levels of overtoxic recommendations in our simulation study, the trade-off between increased ability to select and assign patients to the MTD and potential safety loss is in favor of the calibrated version of the MVN prior. In cases where safety is of more concern, the calibration of the MVN prior parameters could also be done in a two-stage way, ensuring the optimization of both performance and safety [3].

The primary contrast between the power prior and robust MAP approaches is that the former employs static borrowing, with pre-defined exponents, while the latter employs dynamic borrowing. As a consequence, the power prior is unable to accommodate unforeseen disparities between historical and current data, and it is necessary to maintain the same level of confidence in the historical data regardless of the present outcome. As seen in this study, this behaviour of the power prior results in a low percentage of correct selections when the MTD is higher in the human data than in the preclinical data, and high overtoxic recommendations and assigments when the MTD is lower. When downweighting the power prior, both the prior ESS and operating characteristics tend to that of a MVN prior, which shows better performance in case of discordance. The analysis of other extensions of the power prior, such as the modified power prior, is possible. However, these methods are known to require extensive computing [14, 23, 12], have the potential to excessively attenuate the influence of historical data [36], and may require highly informative distributions for the exponential parameters. Therefore, where there is a conflict between prior and current data, the robust MAP approach is expected to outperform the power prior approach, as confirmed in this work. In contrast, the power prior is generally anticipated to exhibit lower variance compared to the robust MAP approach [35], since the robustification process substantially boosts the overall variance of the distribution and the quantity of information contained in the tails. As a result, if there is agreement between the historical and current data, the power prior might outperform the robust MAP approach, again confirmed by this work. The extent of these outperformances illustrates further the point that the MAP approach is more robust, as the losses of the power prior far outweigh the losses of the MAP prior. In the safer scenarios, the power priors show unacceptably conservative

escalation behaviour, with very few patients assigned to the MTD.

One of the limitations of this work is perhaps the reliance on allometric scaling to convert animal doses to human doses. While allometric scaling is appropriate for some drugs, it may not be appropriate for others [33]. However, the use of a random delta parameter helps to account for such cases and evaluate the suitability of the scaling method.

The study conducted by Olson et al. did not impose any restrictions on the time frame for submitting qualifying data sets, and the inclusive years of data collection for the entire database are not disclosed, making it unclear whether the unevenness of study designs over time may have affected the database analysis results. In addition, in cases where a specific toxicity is identified as being most likely dose-limiting, a translational PK/PD approach focused on this aspect may also generate a prior of the dose-response relationship.

The inclusion of a more extensive set of stopping rules was not considered in this work. For example, it may be desirable to stop the trial after a certain number of patients have been assigned to a dose level, given sufficient information is now available on the estimated MTD. It could be useful to explore such rules further, especially given the results presented here. It may also be of interest to explore the incorporation of animal data into model-assisted designs to assess the added benefits, given the flexibility of these methods.

In this work, we aimed to address a specific question: evaluating the contribution of the available preclinical data to the design of an upcoming phase I dose-escalation trial for a specific project. As preclinical data is designed to inform the potential human response to the new compound, there should always be a certain level of confidence that the available preclinical data could be used to inform the dose escalation strategy, and not only the starting dose. However, translating pre-clinical findings to humans requires careful modeling and parametrization. Our approach demonstrates how pre-clinical data can be effectively utilized to guide dose-escalation decisions while ensuring a well-balanced integration of animal and human data. Indeed, the external concordance data that our novel methodology uses to define the parameters of the prior is one more way to ensure consistency with previously observed data, and the robust component ensures protection in case of discordance between human and animal data. Surveys

on concordance ratios between animal and human and depending on therapeutic classes like the one by Olson et al. [27] provide useful information to guide the use of preclinical data depending on the species used in preclinical data and the therapeutic class. The results highlight that pre-clinical data can indeed provide value ensuring the trial design aligns with ethical and safety standards.

In the context of our research, and given the toxicity profile of the animal data, the explored models with the different techniques to integrate pre-clinical information provide reasonable operating characteristics with a safeguard strategy for toxic scenarios.

There is a high interest in improving dose finding strategies by including preclinical data in dose finding studies, as highlighted by the recommendation of the Optimus project initiated by the FDA Oncology Center of Excellence to "develop strategies for dose finding and dose optimization that leverages nonclinical and clinical data in dose selection". In this context, this work serves as a solid starting point and highlights the potential of integrating preclinical data into early-phase clinical trial designs. It also provides motivation to pursue further research in this direction. Expanding our simulations to include both preclinical and clinical data would allow us to explore a broader range of scenarios and better understand the generalizable benefits of incorporating preclinical data into phase I designs. Such an approach could provide deeper insights into the contexts where preclinical data are most impactful, ultimately leading to more informed and effective trial designs.

Acknowledgments

This research is supported by Institut de Recherches Internationales Servier.

Author contributions

Helen Barnett conducted the work and executed the simulation studies. Marie-Karelle Riviere supervised the project and performed double validation. Pavel Mozgunov and Gaëlle Saint-Hilary provided guidance on the research directions and the approaches to be explored and

included in the comparison. Donia Skanji and Sandrine Guilleminot provided the research question and the context for the motivating trial. Fulvio Di Stefano supported the preliminary research work. Mélanie Guhl implemented the revisions in the reviewing process.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

Supporting information

All data is simulated as outlined in the manuscript.

Online Supplementary Materials, including prior calibration, detailed simulation results and sensitivity analysis are available online.

References

[1] James Babb, André Rogatko, and Shelemyahu Zacks. Cancer phase i clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine*, 17(10):1103–1120, may 1998. DOI: 10.1002/(sici)1097-0258(19980530)17:10¡1103::aid-sim793¿3.0.co;2-9.

[2] Helen Barnett, Oliver Boix, Dimitris Kontos, and Thomas Jaki. Joint tite-crm: A design for dose finding studies for therapies with late-onset safety and activity outcomes. *Statistics in Biopharmaceutical Research*, pages 1–12, 2024.

[3] Helen Barnett, Matthew George, Donia Skanji, Gaelle Saint-Hilary, Thomas Jaki, and Pavel Mozgunov. A comparison of model-free phase i dose escalation designs for dual-agent combination therapies. *Statistical Methods in Medical Research*, 33(2):203–226, 2024.

[4] Bradley Carlin and Thomas Louis. *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*. Chapman and Hall/CRC, June 2000. DOI: 10.1177/1740774510382799.

[5] Ming-Hui Chen and Joseph G. Ibrahim. Power prior distributions for regression models. *Statistical Science*, 15(1), feb 2000. DOI: 10.1214/ss/1009212673.

[6] Ming-Hui Chen and Joseph G. Ibrahim. The relationship between the power prior and hierarchical models. *Bayesian Analysis*, 1(3), September 2006. DOI: 10.1214/06-ba118.

[7] Ying Kuen Cheung. *Dose finding by the continual reassessment method*. CRC Press, 2011.

[8] European Medicines Agency. Ich guideline s9 on nonclinical evaluation for anticancer pharmaceuticals. December 2008.

[9] European Medicines Agency. Repeated dose toxicity - Scientific guideline. `https://www.ema.europa.eu/en/repeated-dose-toxicity-scientific-guideline`, 2010.

[10] European Medicines Agency. ICH M3 (R2) Non-clinical safety studies for the conduct of human clinical trials for pharmaceuticals - Scientific guideline. `https://www.ema.europa.eu/en/ich-m3-r2-non-clinical-safety-studies-conduct-human-clinical-trials-pharmaceuticals-`
2013.

[11] Food and Drug Administration. Estimating the maximum safe starting dose in initial clinical trials for therapeutics in adult healthy volunteers - guidance for industry. July 2005.

[12] Byron J. Gajewski. Comments on 'a note on the power prior'. *Statistics in Medicine*, 29(6):708–709, feb 2010. DOI: 10.1002/sim.3824.

[13] Isaac Gravestock and Leonhard Held and. Adaptive power priors with empirical bayes for clinical trials. *Pharmaceutical Statistics*, 16(5):349–360, jun 2017. DOI: 10.1002/pst.1814.

[14] Joseph G. Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in Medicine*, 34(28):3724–3749, sep 2015. DOI:10.1002/sim.6728.

[15] Thomas Jaki, Sally Clive, and Christopher J. Weir. Principles of dose finding studies in cancer: a comparison of trial designs. *Cancer Chemotherapy and Pharmacology*, 71(5):1107–1114, January 2013. DOI: 10.1007/s00280-012-2059-8.

[16] Christophe Le Tourneau, J. Jack Lee, and Lillian L. Siu. Dose Escalation Methods in Phase I Cancer Clinical Trials. *JNCI: Journal of the National Cancer Institute*, 101(10):708–720, 05 2009.

[17] Jessica Lim, Rosalind Walley, Jiacheng Yuan, Jeen Liu, Abhishek Dabral, Nicky Best, Andrew Grieve, Lisa Hampson, Josephine Wolfram, Phil Woodward, Florence Yong, Xiang Zhang, and Ed Bowen. Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: Review of methods and opportunities. *Therapeutic Innovation & Regulatory Science*, 52(5):546–559, sep 2018. 10.1177/2168479018778282.

[18] Nina Magnolo, Külli Kingo, Vivian Laquer, John Browning, Adam Reich, Jacek C. Szepietowski, Deborah Keefe, Philemon Papanastasiou, Weibin Bao, Pascal Forrer, and Manmath Patekar. Efficacy of secukinumab across subgroups and overall safety in pediatric patients with moderate to severe plaque psoriasis: Week 52 results from a phase III randomized study. *Pediatric Drugs*, 24(4):377–387, June 2022. DOI: 10.1007/s40272-022-00507-0.

[19] Satoshi Morita, Peter F Thall, and Peter Müller. Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602, jun 2008.

[20] Pavel Mozgunov, Thomas Jaki, Ioannis Gounaris, Thomas Goddemeier, Anja Victor, and Marianna Grinberg. Practical implementation of the partial ordering continual reassessment method in a phase i combination-schedule dose-finding trial. *Statistics in Medicine*, 41(30):5789–5809, 2022.

[21] Timothy Mutsvari, Dominique Tytgat, and Rosalind Walley. Addressing potential prior-data conflict when using informative priors in proof-of-concept studies. *Pharmaceutical Statistics*, 15(1):28–36, nov 2015. DOI: 10.1002/pst.1722.

[22] Beat Neuenschwander, Michael Branson, and Thomas Gsponer. Critical aspects of the bayesian approach to phase i cancer trials. *Statistics in Medicine*, 27(13):2420–2439, 2008. DOI:10.1002/sim.3230.

[23] Beat Neuenschwander, Michael Branson, and David J. Spiegelhalter. A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566, sep 2009. DOI: 10.1002/sim.3722.

[24] Beat Neuenschwander, Gorana Capkun-Niggli, Michael Branson, and David J Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18, jan 2010. DOI: 10.1177/1740774509356002.

[25] Lei Nie, Eric H. Rubin, Nitin Mehrotra, José Pinheiro, Laura L. Fernandes, Amit Roy, Stuart Bailey, and Dinesh P. de Alwis. Rendering the 3 + 3 design to rest: More efficient approaches to oncology dose-finding trials in the era of targeted therapy. *Clinical Cancer Research*, 22(11):2623–2629, May 2016. DOI: 10.1158/1078-0432.ccr-15-2644.

[26] Adrien Ollier, Satoshi Morita, Moreno Ursino, and Sarah Zohar. An adaptive power prior for sequential clinical trials – application to bridging studies. *Statistical Methods in Medical Research*, 29(8):2282–2294, nov 2019. DOI: 10.1177/0962280219886609.

[27] Harry Olson, Graham Betton, Denise Robinson, Karluss Thomas, Alastair Monro, Gerald Kolaja, Patrick Lilly, James Sanders, Glenn Sipes, William Bracken, Michael Dorato, Koen Van Deun, Peter Smith, Bruce Berger, and Allen Heller. Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regulatory Toxicology and Pharmacology*, 32(1):56–67, aug 2000. DOI: 10.1006/rtph.2000.1399.

[28] John O'quigley, Xavier Paoletti, and Jean Maccario. Non-parametric optimal design in dose finding studies. *Biostatistics*, 3(1):51–56, 03 2002.

[29] John O'Quigley, Margaret Pepe, and Lloyd Fisher. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, pages 33–48, 1990.

[30] Luca Richeldi, Arata Azuma, Vincent Cottin, Christian Hesslinger, Susanne Stowasser, Claudia Valenzuela, Marlies S. Wijsenbeek, Donald F. Zoz, Florian Voss, and Toby M. Maher. Trial of a preferential phosphodiesterase 4b inhibitor for idiopathic pulmonary fibrosis. *New England Journal of Medicine*, 386(23):2178–2187, June 2022. DOI: 10.1056/nejmoa2201737.

[31] Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, oct 2014. DOI: 10.1111/biom.12242.

[32] Heinz Schmidli, Beat Neuenschwander, and Tim Friede. Meta-analytic-predictive use of historical variance data for the design and analysis of clinical trials. *Computational Statistics & Data Analysis*, 113:100–110, sep 2017. DOI: 10.1016/j.csda.2016.08.007.

[33] Vijay Sharma and John H McNeill. To scale or not to scale: the principles of dose extrapolation. *British Journal of Pharmacology*, 157(6):907–921, July 2009. DOI: 10.1111/j.1476-5381.2009.00267.x.

[34] U.S. Food & Drug Administration. M3(R2) Nonclinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorization for Pharmaceuticals. `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/m3r2-nonclinical-safety-studies-conduct-human-clinical-trials-and-marketing-authoriza` 2019.

[35] Joost van Rosmalen, David Dejardin, Yvette van Norden, Bob Löwenberg, and Emmanuel Lesaffre. Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research*, 27(10):3167–3182, feb 2017. DOI: 10.1177/0962280217694506.

[36] Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G. Ibrahim, Nelson Kinnersley, Stacy Lindborg, Sandrine Micallef, Satrajit Roychoudhury, and Laura Thompson. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54, aug 2013. DOI: 10.1002/pst.1589.

[37] Haiyan Zheng, Lisa V Hampson, and Simon Wandel. A robust bayesian meta-analytic approach to incorporate animal data into phase i oncology trials. *Statistical Methods in Medical Research*, 29(1):94–110, jan 2019. DOI: 10.1177/0962280218820040.

[38] Heng Zhou, Thomas A Murray, Haitao Pan, and Ying Yuan. Comparative review of novel model-assisted designs for phase i clinical trials. *Statistics in Medicine*, 37(14):2208–2222, 2018.

[39] Heng Zhou, Ying Yuan, and Lei Nie. Accuracy, safety, and reliability of novel phase i trial designs. *Clinical Cancer Research*, 24(18):4357–4364, September 2018. DOI: 10.1158/1078-0432.ccr-18-0168.