BAYESIAN ADAPTIVE METHODS TO INCORPORATE PRECLINICAL DATA INTO PHASE I CLINICAL TRIALS

Haiyan Zheng

DISSERTATION

Submitted for the degree of
Doctor of Philosophy
at Lancaster University

Medical and Pharmaceutical Statistics Research Unit
Department of Mathematics and Statistics, Lancaster University, U.K.

November 2018

Date of submission: 09.11.2018
PhD viva date: 09.01.2019

Supervisor: Prof. Dr. Thomas Jaki (Lancaster University)
Dr. Lisa V. Hampson (Novartis Pharma AG)

To my parents, without whom none of this would have been possible.

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not ever been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 55,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 40 figures. All research work presented in this dissertation has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 633567. Main results writen in Chapter 3 have been published in *Statistical Methods in Medical Research*.

Haiyan Zheng
November 2018

ACKNOWLEDGEMENTS

The fabulous journey started from three years ago, when I was selected by IDEAS (a Marie Skłodowska-Curie Initial Training Network, `http://www.ideas-itn.eu/`) as one of the fourteen early stage research fellows. I have never imagined pursuing a PhD could be such fun, but IDEAS has tremendously enriched my experience in every aspect. Three years have gone so fast; what remains are fruitful collaborations and long-standing friendship.

First and foremost I would like to thank my supervisors Dr Lisa Hampson and Professor Thomas Jaki, who have set great examples of perfectionism as an excellent researcher and wonderful mentor. Their guidance and constant encouragement have helped me overcome obstacles along the way and eventually become an independent researcher. The supervision meetings were no doubt extremely helpful and enjoyable. I am eternally grateful to Lisa, who trained me well to solve problems as a Bayesian and has been so kind to continue working with me even after her leaving Lancaster University. I was very much impressed by her detailed corrections and suggestions throughout the write-up of our papers. I also owe a large debt of gratitude to Thomas for giving me freedom to explore methodology issues of my central interest and for his more than generous contribution of time and insightful comments. I have been benefited from every conversation with him. He taught me to think independently and work collaboratively, and has always been willing to help connect the dots when things do not seem to line up.

I am also much appreciative of support and insights from Dr Beat Neuenschwander and Dr Simon Wandel based in Novartis. I met them for the first time at the ISCB conference in Birmingham 2016. They kindly arranged logistics for me to spend a secondment with the Methodology Group. This has led to an exciting joint research, of which the results were presented in Chapter 3 of this thesis. My visit to Novartis has been indispensable also for that I learnt to communicate effectively with non-statistician scientists. Our joint work was benefited from discussions with Dr Michéle Bouisset-Leonard, who patiently explained the principles of practical translational science to me. Thanks also go to Dr Amy Racine and Dr Heinz Schmidli for their interest in my PhD research. I particularly enjoyed having their thoughtful input, which indeed helped stimulate more interesting research questions to work on. I would also like to thank others who have helped improve my understanding in early drug development including Dr Alun Bedding who invited me for a research visit in AstraZeneca, and Dr Tom Jacobs who encouraged me to develop sophisticated pharmacological models to incorporate animal data into a first-in-man trial.

I would like to thank all the friends that I have known from IDEAS and the Maths & Stats department in Lancaster. I am especially thankful to Johanna, who brightened up my life in Europe. She turned me into a runner and better skier. I cherish so much the days when we ran along River Rhine in summer and skied in the Black Forest in winter. I have been much touched by Johanna's constant support and trust. My friendship with Marius has also become stronger during my secondment in Novartis, when he shared his favourites (football games and beers, apart from statistics) with me. Pavel is another great friend, who introduced a wide range of cheeses to me. Camembert, Brie and Stilton were the first three that we had together with wines. Sharing the same office in Lancaster, we had quite a few interesting methodological discussions. I believe we will have fruitful collaborations in the near future, given our commen research interests. During my visit to Janssen in Belgium, Saswati and Fabiola were my great company. I remember so well that they each kindly invited me for dinner, treating me with the best cuisine from their homeland that I have ever had. It was also great to have Eleni, Nicolás, José and Fabiola visit us in Lancaster for their secondments. A very pleasant trip to Wales was with Saswati and Arsénio; there we learnt how to pronounce a Welsh word like "Cwmcynwyn". I am looking forward to visiting more interesting places together. I am also truly grateful to have known Helen, Matthieu, Matthias, Adam, Julian, Chao, Georg, Robin from our department, and Francesca, Jared, Malen from the Psychology department. We have formed a great hiking club, and I am really desired to go for another hike together! Thanks so much all my dear friends – it has been highly appreciated!

Finally and above all, I would like to reemphasise that none of this would have been possible without ongoing love and support from my family. Papa and mama, you are such wonderful parents raising me to know right from wrong. I am indebted to you for supporting my dreams and ambitions. Being a statistician is the best way that I can see to contribute to the world. Thanks so much for letting this be possible.

Haiyan Zheng
Lancaster and Newcastle
November 2018

ABSTRACT

Basing informed decisions on available, relevant information is essential in all phases of drug development. This is particularly true in early phase clinical trials, when our knowledge about toxicity of a new medicine remains limited. Thus, borrowing of information across seemingly disparate sources is appealing. Statistical literature has been written about augmenting a new clinical trial with data from historical studies designed for similar investigational purpose. But very few has looked into leveraging preclinical data into phase I first-in-man trials.

The work presented in this thesis attempts to fill the gap by providing solutions in the Bayesian paradigm, with purposes of improving the design and analysis of adaptive phase I dose-escalation trials. Specifically, our focus is on the transition step of early drug development, where phase I clinical trials are preceded only with some preclinical information. We see preclinical data as a special type of historical data, say, historical animal data. This is not an obvious application of the existing approaches for data augmentation, since information collected from preclinical studies first needs to be translated to account for potential physiological differences between animals and humans. Furthermore, due to their intrinsic variabilities in drug metabolism, inconsistency between the translated preclinical and clinical data may still emerge however careful and correct the interspecies translation would be completed. We note this thesis will exclusively consider toxicity data, assuming that relationship between dose and risk of toxicity can be adequately described using a two-parameter logistic regression model.

Grounded in Bayesian statistics, our idea is to represent preclinical data into a prior distribution for the dose-toxicity model parameters that underpin the human trial(s). Our aim is to propose robust Bayesian approaches, keeping in mind the possibility that toxicity in humans could be very different from what we have learnt in one or multiple animal species even after appropriate translation. The main challenge in statistical inferences is essentially to address issues of prior-data conflict emerging in a small trial.

This thesis consists of two perspectives on the robust use of preclinical animal data. A "sensible" amount of animal data to be leveraged into the phase I human trial(s) is determined by either (i) assessing the commensurability of the prior predictions

of human toxicity, which are obtained using animal data alone, with the observed toxicity outcomes from the ongoing trial(s), or (ii) fitting a hierarchical model with weakly informative priors placed on the variance parameters. Correspondingly, we have proposed a Bayesian decision-theoretic approach in Chapter 2 and a robust Bayesian hierarchical model in Chapter 3, which build the core of this thesis. We have also extended the Bayesian hierarchical model to address potential heterogeneity between patient groups in Chapter 4, where the methodology has been illustrated in the context of bridging strategies considered in phase I clinical trials planned in various geographic regions.

Throughout, the proposed Bayesian adaptive methods have been elucidated with representative data examples and extensive simulations. Particular attention has been paid to balancing the information from different sources to draw robust inferences. Numerical results show that our proposals have desired properties. More specifically, preclinical data can be essentially discounted when they are in fact inconsistent with the toxicity in humans. In cases of consistency, benefits are seen as increased precision of estimate of the probability of toxicity at a range of doses, and higher proportion of patients allocated to the target dose(s).

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 OVERVIEW OF DRUG DEVELOPMENT

Bringing a new medicine from laboratory experiments eventually to the market is a long-term, complex, and costly process. This requires expertise from many fields such as pharmacology, toxicology, medical science and statistics, for substantial progress to be made. In essence, drug development is a collection of joint effort whereby a promising compound is identified based on extensive basic research in biochemistry, and evaluated in a series of preclinical studies (Rogge and Taft [2016]) and clinical trials (Chow and Liu [2013]). Particularly, within preclinical settings, *in vitro* (within the glass, i.e., in a laboratory environment) and *in vivo* (within a living organism, such as animals) studies need to be performed for preliminary characterisation of toxicity and efficacy profiles, along with the understanding of pharmacokinetics that describes absorption, distribution, metabolisation and elimination of the drug. Preclinical experimental findings pave the way for further evaluation about safety and efficacy of the drug in clinical trials, which are generally classified into four successive phases, say, I, II, III and IV.

When moving to phase I clinical trials, also often termed as first-in-man trials, safety remains a key priority and special concern. The main focus at this stage is to figure out whether the new drug is tolerable in humans and, more specifically, to identify effective yet sufficiently safe dose(s) for evaluation in subsequent studies (Chevret [2006]). The widely accepted maxim is that all therapeutic agents can be toxic in overdosage: a drug is claimed to be safe as long as the risk of toxicity is under a certain tolerance level. With several dose levels given for evaluation, phase I clinical trials are typically designed to estimate the dose associated with a predefined level of pharmacokinetic target or toxicity, involving a small number of human subjects, say, between 20 and 60 healthy volunteers (for relatively non-toxic agents) or patients (for diseases with high mortality). Phase I clinical trials for anticancer therapies usually recruit patients who have failed other prior treatments. The target dose is generally described as the maximum tolerated dose (MTD), defined as a dose that leads to a maximum of certain percentage, for example, 25%, of patients treated with the drug to experience dose-limiting toxicities (DLTs). This particularly makes sense in the oncology setting, where toxicity and clinical effect are assumed to be positively associated, and both increase monotonically with the dose levels. Throughout this thesis, we will discuss drug development in oncology as our main example and use subsequent chapters to describe the methodology motivated in this setting.

A successful phase I oncology trial will lead to a decisive statement about the MTD, which is generally thought of as the highest dose that could be recommended for further evaluations in a subsequent phase II clinical trial (Paoletti and Doussau [2014]). The main interest then becomes to assess the short-term therapeutic effect and continue monitoring severe adverse events due to the drug, running the study on a larger group of patients, most commonly, 100 to 300. If considerable evidence would show that the pharmaceutical product is safe and efficacious, a confirmatory phase III clinical trial will be encouraged to take place in several hundred to over one thousand patients (Friedman et al. [1998]). Of primary interest in phase III trials is whether the drug provides clinical benefit such as increased survival or symptomatic improvement, compared with either an active control (current standard treatment) or placebo, in various clinical settings. With completion of phase I–III trials, findings will be submitted to a regulatory authority, e.g., Food and Drug Administration in the US or the European Medicines Agency in Europe, for marketing authorisation of the drug. A phase IV trial will be launched for post-marketing surveillance on the long-term safety and efficacy.

Despite of recent innovative proposals about combined phase I/II trials (Zohar and Chevret [2007]) as well as seamless phase II/III trials (Stallard and Todd [2011]), preclinical studies and phase I-IV trials are conventionally conducted in isolation. Moreover, data cumulated from previous studies are very rarely incorporated in a formal manner for decision making in a new clinical trial. In this thesis, we aim to bridge this gap by coming up statistical solutions for leveraging external data in a new clinical trial. Majority of our work is devoted to the transition step from preclinical to phase I first-in-man studies. Nevertheless, the proposed methods could be seen as an illustration of what can be achieved in other settings.

1.2 FUNDAMENTALS OF EARLY PHASE CLINICAL TRIALS

The purpose of early phase clinical trials (here, referred to as exploratory studies) typically includes characterising the toxicity profile of a new medicine in humans and establishing effective dosage regimens. In monitoring safety, adverse events will be recorded and graded according to some severity grading scales of adverse events, such as the National Cancer Institute Common Terminology Criteria (National Cancer Institute [2017]). Although a grade 3 (severe) or higher (life-threatening or fatal) toxicity is generally described as dose limiting, definition of DLTs can vary from one study to another; for example prolonged grade 2 toxicities can be considered as DLTs depending on schedule of drug administration. Furthermore, there is no uniformity in the definition of DLTs for molecularly targeted agents, which may display a very distinct toxicity profile following continuous and prolonged administration compared with cytotoxic agents (Le Tourneau et al. [2011]; Bautista et al. [2017]). In order to reach timely decision making in the dose recommendation process, either a DLT or no DLT will be summarised for each patient by the end of the first treatment cycle. Care needs to be taken for trials where possibility of late-onset toxicities is an important concern.

For the simplified scenario with a binary safety endpoint, the MTD can be regarded as a quantile of a dose-toxicity probability curve. Let Y_j be the dichotomous outcome of a patient, who experiences either a DLT ($Y_j = 1$) or no DLT ($Y_j = 0$) on receiving a dose d_j chosen from a predefined discrete dosing set $\mathcal{D} = \{d_1, \dots, d_J\}$. We note this is a tool of convenience, and are aware of situations where continuous doses are used (Diniz et al. [2017]). Suppose there have been n_j patients treated per dose d_j. We then know $Y_j \sim \text{Bernoulli}(p_j)$, where p_j is the probability of toxicity at the dose. It is also reasonable to consider probabilities of toxicity at different doses are correlated and can be described using a statistical model $\psi(\cdot)$:

$$p_j = \mathbb{P}(Y_j = 1|d_j) = \psi(d_j), \quad \text{for } y_j = 0, \dots, n_j.$$

On the termination of a phase I trial, the MTD is declared based on probabilistic inference about the toxicity rate p_j implied by the dose-toxicity model $\psi(d_j)$, which describes the randomness of toxicity outcome after administration of a dose. For a predefined target level of Γ, the MTD, denoted by d_M, can be estimated as

$$d_M = \psi^{-1}(\Gamma),$$

where the $\psi^{-1}(\cdot)$ is an inverse of the dose-toxicity model. Indeed, this is regarded as an estimation problem rather than statistical testing of a hypothesis. Different models can be considered, depending on many aspects including experiment designs, knowledge of underlying biological mechanism, and possible stochastic effects such as random errors and "population" variability. Of note, phase I studies are not limited to the "first-in-man" trials, but there could be subsequent phase Ib studies to evaluate new administration schedules or combinations of established agents (Weber et al. [2015]). We will illustrate the key statistical properties of dose-escalation designs for phase I oncology trials in Section 1.2.1. For keeping it simple, we will restrict our focus to first-in-man trials for monotherapy in the following.

1.2.1 *Design and analysis of dose-escalation trials*

In 2013, the World Medical Association released guidelines on ethical considerations for medical research involving human subjects (World Medical Association [2013]). Investigators conducting clinical trials must adhere to the ethical norms, despite that researcher' goals may differ from those of patients'. A good experimental design for phase I clinical trials is one that supports efficient estimation of quantiles of interest without exposing many patients to doses that are overly toxic. This means random allocation of patients to doses contained in $\mathcal{D} = \{d_1, \dots, d_J\}$ is unacceptable (or more accurately, unethical) unless all the doses in set $\mathcal{D}$ have been proven to have similar toxicity profiles and are potentially efficacious to different extent. However, sufficient toxicity data in humans must be accumulated to verify this presumption which may not always be true. To facilitate the decision making process with the most up-to-date knowledge about toxicity, phase I clinical trials are often designed in an adaptive way with specification of

- a safe starting dose, commonly, the lowest dose $d_1 \in \mathcal{D}$,
- sequential accrual of patients to be treated in small cohorts until reaching the maximum sample size,
- criteria for interim dose escalation and de-escalation,
- additional stopping rules.

Issues raised in determining a safe starting dose will be discussed in Section 1.3, while this section will review several classes of dose-escalation designs developed assuming doses in $\mathcal{D}$ are sensible and the focus is placed on statistical inference for interim decision making.

Numerous statistical designs have been proposed since 1990s for phase I dose-escalation trials in oncology. There are two divergent schools: algorithmic and model-based methods (Braun [2014]). Although algorithmic methods have been criticised for allocating too many patients to suboptimal doses and giving inaccurate estimate of the MTD, dose-escalation procedures in this class remain a prevailing approach due to the simplicity in logistics for clinical investigators to carry out phase I clinical trials (Le Tourneau et al. [2009]). A diagram for this type of design, termed as "3 + 3" design (Storer [1989]), is shown in Figure 1.1. As we can see, decisions on escalation, termination and declaration of MTD essentially come from data of the latest one or two cohorts, with each constituted by three patients: data collected from previous patient cohorts, especially those who have received a different dose, are completely discarded. This leads to myopic decision making and undesirable under- or over-estimation of toxicity in humans. As evaluated by Lin and Shih [2001], the probability of toxicity at the dose declared as MTD does not converge to a fixed target, say, 33% that may be anticipated by trialists in favour of the 3 + 3 design. Modified algorithmic designs have been proposed in response to the outstanding drawback of the 3 + 3 design; see, for example, a family of accelerated titration dose-escalation designs (Simon et al. [1997]), the biased coin designs proposed by Durham et al. [1997], the group up-and-down designs by Gezmu and Flournoy [2006] and so forth. Despite of the fact that algorithmic designs are implemented by trialists more often in practice, they are not the most efficient ones and are criticised for not using the entire trial history with information that has been collected from all patients treated so far. As an alternative, model-based procedures can facilitate design adaptations for protecting the ethics of phase I first-in-man trials (Love et al. [2017]). We will focus on model-based designs in the following.

In a trial that has evaluated a range of ordered doses, more often than not one may be interested in the underlying dose-toxicity relationship that can be described using, for example, a logistic regression model. Decisions can be made based on the established dose-toxicity model, which is to be updated along with new data accrued from the ongoing trial. The first model-based design for dose-escalation trials is the continual reassessment method (CRM), which addresses practical and ethical concerns in a rigorous mathematical framework (O'Quigley et al. [1990]). Iasonos et al. [2008, 2012] and Jaki et al. [2013] have commented that the CRM has superior operating characteristics to the algorithmic 3 + 3 design. The main idea of the CRM is to sequentially assign incoming patients to a dose, at which the probability of toxicity is closest to the target level, and to update the dose-toxicity relationship with toxicity

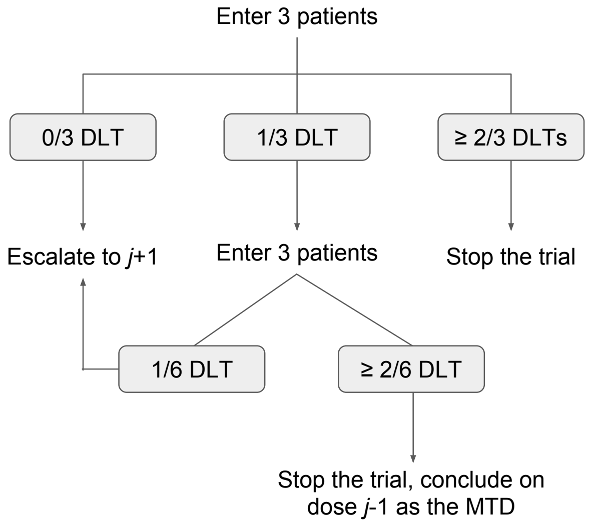

Figure 1.1: Escalation scheme for the traditional 3+3 designs.

outcomes observed from all patients that have been treated so far. As the phase I trial proceeds, a more accurate estimate about the dose-toxicity relationship will be obtained. In other words, current knowledge can be sequentially updated as the new information (here, referred to as the toxicity data from a new patient cohort) comes in. While frequentist approaches can work for this purpose, we will tackle problems in a Bayesian paradigm, where the posterior distribution fomulates naturally a recursive estimator to support a transparent decision making process.

Let $x(i)$ denote the dose chosen from $\mathcal{D} = \{d_1, \dots, d_J\}$ that fulfil certain criteria to treat patient i, and z_i is the binary toxicity outcome observed of this patient. Let $\mathbf{x}_i = \{(x(1), z_1), \dots, (x(i), z_i)\}$ further be the interim data accumulated before we will make a dose recommendation for the $(i+1)$th patient. Typically, a simple parametric mathematical model with a parameter vector θ is assumed. The posterior distribution of θ is derived using Bayes' rule that

$$\pi(\theta|\mathbf{x}_i) = \frac{f(\mathbf{x}_i|\theta)\pi(\theta)}{\int f(\mathbf{x}_i|\theta)\pi(\theta)d\theta}, \tag{1.2.1}$$

where $\pi(\theta)$ is the prior distribution that summarises our preliminary knowledge about the dose-toxicity relationship before the phase I trial starts, and $f(\mathbf{x}_i|\theta)$ is the likelihood function given by

$$f(\mathbf{x}_i|\theta) = \prod_{t=1}^{i} \{\psi(x(t);\theta)\}^{z_t}\{1-\psi(x(t);\theta)\}^{1-z_t}. \tag{1.2.2}$$

Here, we have linked doses $d_j \in \mathcal{D}$ with the corresponding toxicity rates using the assumed mathematical model $\psi(d_j;\theta)$. Selection of dose to be recommended for the next patient cohort may rely on a loss function denoted by $L(\cdot)$ that

$$x(i+1) = \arg\min_{d_k \in \mathcal{D}} L(d_k), \tag{1.2.3}$$

where the loss function can be defined on the scale of probability of toxicity. One widely applied example is the "patient gain criterion", which minimises the distance between the actualised posterior probability of toxicity DLT and the target probability denoted by Γ:

$$L(d_k) = \left| \int \psi(d_k;\theta)\pi(\theta|\mathbf{x}_i)d\theta - \Gamma \right|. \tag{1.2.4}$$

In the field, a number of model-based trial designs, which are conceptually similar, have been proposed since the CRM. These include the designs of escalation with overdose control that adopts an asymmetric loss function to penalise more on overdosing than underdosing by Babb et al. [1998] and the Bayesian logistic regression methods that employs decision theory by Whitehead [2006]. In Figure 1.2, we use a diagram to summarise similarity of the decision making process when applying these Bayesian model-based methods for design and analysis of phase I dose-escalation trials.

Many nonparametric designs have also been developed over the past decades. The distinction between this class of phase I trial designs and the model-based designs described above is that no specific parametric assumptions would be required for the underlying distribution of the toxicity probability in relation to the doses. In the literature, Gasparini and Eisele [2000] (with correction in Gasparini and Eisele [2001]) presented curve-free designs, where they place a multi-dimensional prior directly on the vector of risks of toxicity at all available doses. John et al. [2010] develope a Bayesian procedure that assumes only monotonicity in the dose-toxicity relationship. Other well-known nonparametric proposals include the modified toxicity probability interval (mTPI) method developed by Ji et al. [2010] to recommend a dose based

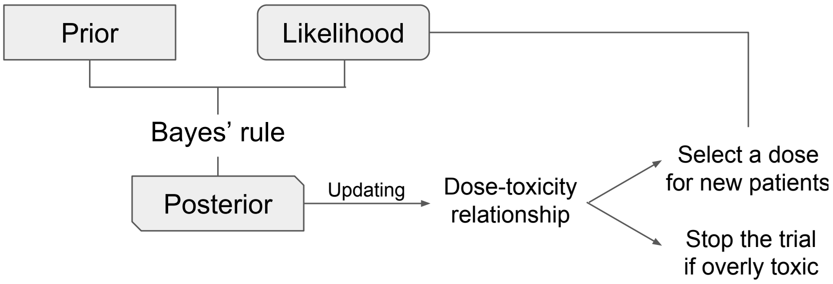
Prior
Likelihood
Bayes' rule
Posterior
Updating
Dose-toxicity relationship
Select a dose for new patients
Stop the trial if overly toxic

Figure 1.2: Escalation scheme for the Bayesian model-based designs.

on unit probability mass, and the Bayesian optimal interval (BOIN) designs by Liu and Yuan [2015] to minimise the probability of incorrect dose selection in a decision-making framework. Horton et al. [2017] compare CRM with mTPI and BOIN in an extensive simulation study, where the evaluation of these methods is undertaken particularly with respect to percentage of correct selection of the true MTD, allocation of patients to doses at or close to the true MTD, and an accuracy index. It was found that CRM outperforms these two alternatives, leading to more efficient and ethical phase I dose-escalation trials, especially when the dose-toxicity curve is characterised with many dose levels, say, six to eight dose levels. These methods present fairly similar behaviours when the number of dose levels for evaluation is decreased.

The operating characteristics of model-based designs and nonparametric designs are comparably good, unless the underlying parametric assumption is substantially incorrect. An evaluation of the properties of the model-based designs and the curve-free designs for phase I dose-escalation trials is presented by Jaki et al. [2013]. We believe in our context it would be beneficial to assume on the form that the entire dose-toxicity curve may take, as the functional model holds promise for prediction of the interpolated and extrapolated data on doses that have not yet been evaluated by the time. We will therefore be concentrated on the model-based approaches in futher discussion about our proposals.

While both frequentist (for example, see O'Quigley and Shen [1996]) and Bayesian model-based approaches to the design and analysis for adaptive phase I clinical trials exist, the Bayesian paradigm offers possibility to integrate prior information, such as the preliminary knowledge about the dose-toxicity relationship learned from external studies. One advantage of increasing the amount of information in a phase I first-in-man trial is that more efficient and ethical decisions on dose assignment can be reached. But this would certainly bear an increased risk of selecting the incorrect dose as the MTD if the prior turns out to be inconsistent with the data. Before we will look into adaptive approaches to discounting inconsistent priors, we first describe two types of prior distributions concerned for dose-toxicity models; specifically, they are operational priors in Chapter 1.2.2 and informative priors in Chapter 1.2.3.

1.2.2 *Operational priors for some parametric dose-toxicity models*

Let us start with the power model commonly considered for the CRM. Denoting the probabilities of toxicity at doses $d_j \in \mathcal{D}$ by p_j, we write the

$$\psi(d_j;\theta) = (p_j)^{\exp(\theta)}, \qquad \text{for } j = 1, \dots, J,$$

where θ is an unknown model parameter. Many authors have chosen to use a normal prior such as $N(0, 1.34)$ for θ (Cheung [2011]), which suggests that with probability 95%, the exponent $\exp(\theta)$ would fall within the interval (0.103, 9.668), together with a set of discrete prior probabilities of toxicity, say, π_j to 'impute' to the toxicity rates p_j. Given the ordered $0 < \pi_1 < \cdots < \pi_J \leqslant 1$, sometimes also termed as 'skeleton' probabilities, large variability is permitted for the toxicity rate per dose, and this mathematical model $\psi(d_j;\theta)$ thus acccommodates flat to very steep dose-toxicity curves.

When considering a two-parameter sigmoid model, $p_j = \psi(d_j;\theta)$ could follow a logistic regression form with model parameters $\theta = (\theta_1, \theta_2)$:

$$\text{logit}(p_j) = \log(\theta_1) + \theta_2 \log(d_j), \quad \theta_1, \theta_2 > 0,$$

following the parameterisation in Neuenschwander et al. [2008]. Investigators may consider a flat improper prior for $\boldsymbol{\theta}$ such that the posterior is proportional to the likelihood. However, this would result in undesirable implications when no DLTs would be observed from the phase I trial (O'Quigley [2002]). Neuenschwander and his colleagues propose to first formulate prior information on the scale of p_j and approximate the priors expressed for p_j by a bivariate normal prior $\pi_0(\boldsymbol{\theta})$. Specifically, J minimally informative beta priors $\text{Beta}(1, a)$ or $\text{Beta}(b, 1)$ with $a \geqslant 1$ or $b \geqslant 1$, will be specified for p_j, with the steps listed as follows.

- Specify two prior quantiles $q(\zeta)$ for toxicity rates p_j at the lowest and highest doses, respectively, such that $\mathbb{P}(p_j < q(\zeta)) = \zeta$;
  - For example, let $q(\zeta) = 0.4$ and $\zeta = 0.95$ for the lowest dose, which means for d_1, with probability 95% (very likely), the toxicity rate p_1 is lower than 40%.
  - Likewise, let $q(\zeta) = 0.2$ and $\zeta = 0.05$ for the highest dose, which means for d_J, with probability 5% (very unlikely), the toxicity rate p_J is lower than 20%.
- Obtain $a = \frac{\ln(1-\zeta)}{\ln(1-q(\zeta))}$ if $q(\zeta) \leqslant \zeta$, or $b = \frac{\ln(\zeta)}{\ln(q(\zeta))}$ if $q(\zeta) > \zeta$;
  - This leads to $p_1 \sim \text{Beta}(1, 5.864)$ and $p_J \sim \text{Beta}(1.861, 1)$ in our example for illustration
- The two prior medians denoted by μ_1 and μ_J are thus known corresponding to the obtained beta prior distributions;
- Assume the prior medians $\mu_1, \ldots, \mu_J$ to be linear in $\log(d_j)$ on the logit scale;
- Substitute the $q(\zeta)$ with the prior medians and let $\zeta = 0.5$, we can obtain other beta priors to describe the toxicity rate at any medium doses d_j.

Whitehead and Williamson [1998] have a similar proposal on specifying priors on the scale of toxicity rate. They imagine any relevant external data can be expressed as information that would have been obtained from a total of six patients to describe the toxicity rates at the lowest and highest doses. Suppose the prior probability of toxicity at the lowest dose is thought to be π_1 and that at the highest dose is π_J. We may consider setting $p_1 \sim \text{Beta}(3 \times \pi_1, 3 \times (1 - \pi_1))$ and $p_J \sim \text{Beta}(3 \times \pi_J, 3 \times (1 - \pi_J))$ for toxicity rates at these two doses, respectively. The joint prior density of p_1 and p_J can be expressed as joint density of θ_1 and θ_2. The theory following this pseudo-observations prior on two doses is mathematically tractable and can be used readily for implementation with standard statistical software.

1.2.3 *Specifying an informative prior using historical data*

The operational priors describe above are not specified using historical data. Instead, they represent opinions directly about the parameter(s) of a new trial, and generally contain least amount of information to advise on plausible values that the new model parameters may take. When there exist sufficient relevant historical data, we may be able to obtain a more informative prior to fit the Bayesian model. With above the examples, the prior distributions could be more informative by having historical data to suggest a small standard deviation in the normal prior for the power parameter, or specify the beta priors with larger effective sample size. However, the trick is to obtain correct link between the historical and new trial data. Several statistical approaches have been proposed to incorporating historical data into a new clinical trial. In the following, we briefly describe how historical data may be represented in a prior for the new parameter(s).

Let $\mathbf{x}_E = \{\mathbf{x}_{01}, \mathbf{x}_{02}, \dots, \mathbf{x}_{0M}\}$ denote the historical data from M existing studies, and $\mathbf{x}_N$ denote the data accumulated from a new trial. With θ_{0i} and θ denoting the trial-specific parameter(s) to underpin either a historical study or a new clinical trial, we let $\mathcal{L}_{0i}(\theta_{0i}|\mathbf{x}_{0i})$ and $\mathcal{L}(\theta|\mathbf{x}_N)$ be the likelihood functions correspondingly.

- Pocock's approach to account for bias
  The main idea of the approach proposed by Pocock [1976] is to model the bias between each historical parameter θ_{0i} and the new trial parameter θ. More, specifically, a bias parameter is defined as a random variable $\delta_i = \theta - \theta_{0i}, i = 1, \dots, M$ and that $\delta_i \sim N(0, \sigma_\delta^2)$, where σ_δ^2 (commonly assumed to be known) suggests the amount of between trial heterogeneity. When the new trial data become available, the posterior distribution can be given by

  $$\pi^P(\theta, \delta_1, \dots, \delta_M | \sigma_\delta^2, \mathbf{x}_E, \mathbf{x}_N) \propto L(\theta|\mathbf{x}_N) \left(\prod_{i-1}^{M} f(\delta_i | 0, \sigma_\delta^2) \mathcal{L}_{0i}(\theta - \delta_i | \mathbf{x}_{0i}) \right) \pi(\theta),$$

  where $f(\cdot)$ is the probability density function of a normal variable, and $\pi(\theta)$ is an uninformative prior for the new trial parameter(s).

- Power prior and modified power prior
  Ibrahim and Chen [2000] propose to raise the likelihood of the historical data to a power $a_{0i} \in [0, 1]$, specific to each historical study, defining that

  $$\pi^{PP}(\theta, a_{01}, \dots, a_{0M} | \gamma_0, \mathbf{x}_E) \propto \left(\prod_{i-1}^{M} \mathcal{L}_{0i}(\theta|\mathbf{x}_{0i})^{a_{0i}} \pi_0(a_{0i}|\gamma_0) \right) \pi(\theta),$$

  where a common set of parameters θ for historical and new trial data has been assumed, and γ represents the hyperparameters for the discount parameters

a_{0i}. The authors suggest to place a beta or truncated gamma prior on a_{0i}. With inclusion of new trial data, the posterior yielded by power prior would be

$$\pi^{\mathrm{PP}}(\theta, a_{01}, \dots, a_{0M} | \gamma_0, \mathbf{x}_E, \mathbf{x}_N) \propto \mathcal{L}(\theta|\mathbf{x}_N) \left(\prod_{i-1}^{M} \mathcal{L}_{0i}(\theta|\mathbf{x}_{0i})^{a_{0i}} \pi_0(a_{0i}|\gamma_0) \right) \pi(\theta).$$

Duan et al. [2006] and Neuenschwander et al. [2009] noted the initial version of the power prior $\pi^{\mathrm{PP}}(\cdot)$ violates the likelihood principle, and proposed adding a normalising constant $C(a_0) = \int_\theta \left(\prod_{i-1}^{M} \mathcal{L}_{0i}(\theta|\mathbf{x}_{0i})^{a_{0i}} \pi_0(a_{0i}|\gamma_0) \right) \pi(\theta) d\theta$:

$$\pi^{\mathrm{MPP}}(\theta, a_{01}, \dots, a_{0M} | \gamma_0, \mathbf{x}_E) \propto \frac{1}{C(a_0)} \left(\prod_{i-1}^{M} \mathcal{L}_{0i}(\theta|\mathbf{x}_{0i})^{a_{0i}} \pi_0(a_{0i}|\gamma_0) \right) \pi(\theta),$$

which is later referred to as the modified power prior.

- Commensurate prior
  Hobbs et al. [2011] develop the commensurate prior to parameterise explicitly on the between-trial heterogeneity for cases where there is only one historical dataset $\mathbf{x}_E = \mathbf{x}_{01}$. We will use θ_0 to denote the single historical parameter here. Following the proposal of commensurate prior, $\theta|\theta_0, \sigma \sim N(\theta_0, \sigma^2)$, where the variance σ^2 controls the degree of borrowing across trials. The commensurate prior is conceptually very similar to Pocock's approach. A conditional prior for θ can therefore obtained:

  $$\pi^{\mathrm{CP}}(\theta|\mathbf{x}_E, \sigma) \propto \int_{\theta_0} f(\theta|\theta_0, \sigma)\pi_0(\sigma)\pi_0(\theta_0)\mathcal{L}_0(\theta_0|\mathbf{x}_E) d\theta_0,$$

  where $f(\cdot)$ is the probability density function of a normal variable and $\pi_0(\sigma)$ is an uninformative prior placed on σ.

- Meta-analytic approach and the robust version
  Neuenschwander et al. [2010] consider to leverage historical data into a new clinical trial based on Bayesian random-effects meta-analysis, assuming that the historical parameters $\theta_{01}, \dots, \theta_{0M}$ are exchangeable with the new parameter θ

under a normal distribution with unknown mean μ and unknown variance τ. Formally, we would write, for $i = 1, \dots, M$,

$$\begin{aligned} X_{0i}|\theta_{0i} &\sim N(\theta_{0i}, s_i^2) \\ \theta_{0i}|\mu, \tau &\sim N(\mu, \tau), \\ \mu &\sim \pi_1(\mu), \\ \tau &\sim \pi_2(\tau), \end{aligned}$$

where X_{0i} denote the outcome of interest in the historical dataset $\boldsymbol{x}_{01}$ and s_i^2 are known variances for the normally distributed historical estimates. Furthermore, for the new trial parameter, we assume

$$\theta|\mu, \tau \sim N(\mu, \tau).$$

When $\boldsymbol{x}_N$ become available, the meta-analytic posterior is given by

$$\pi^{MA}(\theta|\mu, \tau, \boldsymbol{x}_E, \boldsymbol{x}_N) \propto \mathcal{L}(\theta|\boldsymbol{x}_N)\left(\prod_{i=1}^{M} f(\theta_{0i}|\mu, \tau)\mathcal{L}_{0i}(X_{0i}|\boldsymbol{x}_{0i})\right)\pi(\mu)\pi(\tau).$$

Schmidli et al. [2014] propose a robust version of the meta-analytic prior by including a weakly informative prior π_0^R to account for probability of non-exchangeability:

$$\pi^{RMA}(\theta|\mu, \tau, \boldsymbol{x}_E) = w \times \pi^{MA}(\theta|\mu, \tau, \boldsymbol{x}_E) + (1 - w) \times \pi_0^R,$$

where w is the prior probability that the new trial parameter θ is exchangeable with its historical counterpart, $\theta_{01}, \dots, \theta_{0M}$.

1.3 USING PRECLINICAL ANIMAL DATA: CHALLENGES AND OPPORTUNITIES

There are strong scientific and ethical arguments for preliminarily characterising the toxicity profile of a new compound in animals before it will be evaluated in human subjects. A typical preclinical development plan often consists of (i) *in vitro* assays for the identification of an active chemical compound, (ii) *in vivo* studies to assess potential antitumor activity; (iii) toxicology studies to characterise toxicity in animals of at least two species, say, one rodent and the other non-rodent, and (iv) pharmacological studies to elucidate mechanism of the drug action. By definition, (iii) and

(iv) are *in vivo* laboratory animal testing. Both can help extrapolate doses, especially a safe starting dose, for further evaluation of the drug in a first-in-man trial. These two types of preclinical research are of the most interest in the present thesis topic.

Ideally, preclinical data collected from well-planned animal studies will be used for predicting a therapeutic range to focus on in the subsequent clinical testing, wishing that the first tests of drug in humans can be reasonably safe. Several guidelines and points-to-consider documentations have been issued by regulatory agencies to state the importance of preclinical safety evaluation to support clinical drug development (EMA [2005a, 2009, 2011]; FDA [2013]). We give an overview about how preclinical animal data are used in early drug development in Chapter 1.3.2, and discuss what could be possibly achieved beyond this in Chapter 1.3.3.

1.3.1 *Considerations of animal data for extrapolation*

Recent reports (Roberts et al. [2002]; Hackam and Redelmeier [2006]; Hirst et al. [2015]; Macleod et al. [2015]) have called into questions on the reproducibility and translatability of preclinical animal studies. One set of criticism surrounds the need to improve experimental design, conduct and statistical analysis of preclinical research. Editorial Office of the British Journal of Pharmacology established new guidance (Curtis et al. [2015]), where a number of important issues relating to the planning and reporting of animal experiments have been highlighted. The most fundamental consideration may be the sample size, i.e., number of animals, required to provide adequate amount of information about the safety and efficacy of the drug. Power analysis (Festing and Altman [2002]) is favoured as a scientific approach for sample size calculation. This would require the investigators to assume a desired effect size and standard deviation, together with a nominal α level of significance and the power $1 - \beta$ of a specified statistical test, which depends on the objectives of the study in question. For instance, in a screening experiment that aims to declare superiority of an active dose versus placebo or a very low dose in terms of the response rates, a chi-squared statistic or Fisher's exact test may be suggested; when multiple treatment groups are involved, it would require tests for the linear trend that can be based on large-sample chi-square statistic or on exact permutation tests. In practice, logistical and budgetary constraints also serves to limit the size of animal experiments.

A second fundamental aspect is the allocation of available animals to various groups of interest in possibly the best way. Investigators could assign animals to each dose group equally, which may probably work for most cases. But such balanced allocation might not be the most optimal strategy. Relating to the allocation, a vital concern is proper randomisation, which permits fair and valid statistical inferences to be based on the probability of any observed effects owing to chance alone. To draw a conclusion that any observed differences between groups of comparison are due to the treatments themselves rather than to other intrinsic between-group variability, additional assumptions may be required for animal experiments without randomisation. For example, bias would be introduced if animals allocated to one group are more likely to develop tumours than those in another group, but randomisation of animals to each group can protect against such bias. Bias could also be induced on a consciousness level, say, the observer bias, when conducting an animal experiment, which can be coped with by proper blinding.

A recent review of 2 000 published preclinical research showed that these important statistical design aspects have received very scant attention (Macleod et al. [2015]). The Editorial Office of the British Journal of Pharmacology has recently established a new guidance setting up the standard for animal experimental design, analysis and reporting (Curtis et al. [2015]). Hooijmans et al. [2018] presented a GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) approach to rate the certainty in the preclinical evidence that can be used to inform decisions to be made during a clinical research in humans. The GRADE framework has particularly valued the specification of patient relevant outcomes, study limitations, risk of bias, and precision of results (Balshem et al. [2011]; Guyatt et al. [2011c,a,b]). In chapters where we will describe our proposals, we assume the preclinical animal data have been carefully selected regarding the quality and reliability.

In practice, the most common scenario for preclinical animal studies remains to be a comparison between groups (usually less than three) for assessing potential differences in toxicity or effect of a medicine on a qualitative basis. Sophisticated animal studies describing the characteristics of a dose-toxicity curve, or even more specific, a dose-exposure-toxicity curve, in quantitative terms are encouraged to be established.

1.3.2 *Establishing a safe starting dose*

It is crucial to the success of a first-in-man trial that a safe but sensible starting dose can be determined from a preclinical package. Starting with a high dose can result in immediate toxicity and early termination of a phase I trial, while a too low dose would add unnecessary extra testing and also incur ethical concerns that patients may be treated with sub-optimal doses. FDA draft guideline *Estimating the Maximum Safe Starting Dose in Initial Clinical Trials for Therapeutics in Adult Healthy Volunteers* advised on a stepwise algorithm of using preclinical toxicology data to establish a safe starting dose in first-in-man trials USFDA [2005]. Considerations include selection of the most appropriate animal species, determination of the no observed adverse effect levels (NOAELs) in the tested animal species, conversion of NOAELs to human equivalent doses (HED), and application of a safety factor.

However, there is no "one size fits all" approach to select the most relevant animal species. Rather, it must be tailored to the specific investigational agent and meantime requires the input of subject-matter experts such as translational scientists or pharmacologists. A particular species is claimed to be appropriate if it provides good predictability of human toxicity risk, rather than being the most sensitive animal species to the drug. One necessary but not sufficient condition of achieving satisfactory predictability is that the DLTs of a given drug in animals are consistent with those expected in humans. Any preclinical data on the DLTs specific only to the tested animal species but not to humans should be discarded, as they are of limited value for human toxicity risk assessment. In situations when all tested animal species predict comparably similar toxicity in humans or when no further information is available to aid the selection of animal species, the most sensitive species may be selected to gauge a most conservative starting dose in human trials.

After preclinical data from the most relevant animal species are made available, a NOAEL (usually reported in mg/kg) can be determined for each animal study. Such NOAELs refers to the highest dose level that does not produce significant side effects compared with the control group. These will then be converted to HEDs: animal doses are assumed to be scaled well between species when they are normalised to body surface area. Following the FDA guideline, we have

$$\text{HED (in mg/kg)} = \text{animal dose (in mg/kg)} \times \frac{(\text{BW}/\text{BSA})_{\text{Animal}}}{(\text{BW}/\text{BSA})_{\text{Human}}},$$

where BW denotes the body weight (in kg) and BSA is the body surface area measured in square metres; see Sharma and McNeill Sharma and McNeill [2009] for further details and principles of allometric scaling of doses across species. A safety factor is recommended to be applied to the calculated HED from the most relevant animal species. This is deemed to provide a marging of safety to address additional variability between animal species and humans, avoiding overly toxic initial clinical dose.

1.3.3 *Incorporating the whole toxicity profile*

Current approaches to using preclinical animal data culminate in a safe starting dose for a phase I first-in-man trial. This underutilises the toxicity data accumulated from the animal studies. We see preclinical animal data as a special type of historical data for a phase I clinical trial designed to evaluate the same drug in human subjects. A neat way to incorporate the whole toxicity profile characterising the animal data is indeed to summarise such information with a dose-toxicity model parameter (vector), say, $\theta_{\mathcal{A}}$. Likewise, we would have a separate dose-toxicity model parameter (vector) to underpin the phase I human trial, denoted by $\theta_{\mathcal{H}}$. We have briefly reviewed several existing statistical approaches in Chapter 1.2.3 to associating the new trial parameter with historical parameters. However, since the dose-toxicity curves learnt from animal studies may be defined in a dose region that is completely inapproapriate for humans, it is not suitable to substitute the historical parameters in Chapter 1.2.3 with $\theta_{\mathcal{A}}$.

This motivates our investigation of the feasibility of Bayesian adaptive methods for leveraging animal data on toxicity into phase I clinical trials. Challenges mainly include how to proper address the (i) uncertainty that surrounds the current practice in translating animal data to an equivalent human dosing scale, and (ii) possibility that inconsistency between preclinical and clinical data could occur even after very careful selection of animal data and correct interspecies translation.

1.4 thesis organisation

The remainder of the thesis evolves as follows. In Chapter 2, we propose a Bayesian decision-theoretic approach to adaptively incorporiate preclinical animal data, which are captured as an informative component of a mixture prior for the dose-toxicity parameters that underpin the phase I first-in-man trial. Particularly, we assess the prior mixture weight allocated to the preclinical data prior dynamically as the trial progresses, based on how commensurate the prior predictions of human responses would be with the actual observations. Such prior predictions are optimal in the sense of maximising the prior expected utility, by assigning correct predictions a utility of 1, incorrect predictions a utility of 0, or a utility between 0 and 1, depending on whether animal data underestimate, or overestimate, the toxicity in humans. The attained predictive utility of the preclinical data, expressed as a fraction of the maximum utility achieved when all observations are correctly predicted, is then used to quantify the prior mixture weight when determining the dose to be recommended for the next patient cohort.

In Chapter 3, we propose a Bayesian hierarchical model to synthesise animal and human toxicity data, using scaling factors to translate the animal doses administered to different species onto an equivalent human dosing scale. Parameters of logistic models for the dose-toxicity relationship in any tested animal species and humans can therefore be interpreted on a common scale. Prior distributions are specified to describe uncertainty about the magnitude of the translation factor appropriate for each species. Within an animal species, the study-specific dose-toxicity parameters are assumed to be exchangeable. Furthermore, the population parameters specific to each animal species, which have also been expressed on the common scale, say, the human-equivalent scale, are assumed to be exchangeable and thus can be modelled with a "supra-species" random-effects distribution to allow for increased borrowing of information between animal species. Finally, robust borrowing of information from animals to humans is permitted by modelling the parameters of a future phase I first-in-man trial as exchangeable with those standardised animal parameters: for each animal species, a prior mixture weight is defined representing our prior scepticism about the plausibility of an exchangeability assumption, while the option of non-exchangeability with animals is also considered. In this way, information is borrowed only from the most relevant animal species.

In Chapter 4, we generalise the methodology proposed in Chapter 3 to supplement phase I dose-escalation trials with co-data, which comprise (i) preclinical animal data from multiple species and (ii) toxicity data from, either completed or ongoing, phase I clinical trials that recruit and treat patients in other geographic regions. We reserve translation parameters in our Bayesian hierarchical model to address the intrinsic differences in toxicity of the drug between animals and humans, and the potential variability across various human subgroups, arising due to differences in genetics, metabolism or factors relating to diet and environment exposure. The human study-specific parameters are then assumed to be drawn from a common exchangeability distribution, where the means are determined by the animal data and the unknown covariance matrix pertinent would explain the extrinsic heterogeneity between the patient ethnic subgroups. Moreover, we permit the possibility of non-exchangeability for each dose-toxicity parameter vector that underpins a phase I clinical trial, to avoid excessive shrinkage for an extreme stratum.

All the Bayesian models newly proposed in this thesis are fitted using Markov chain Monte Carlo and can be implemented with software such as OpenBUGS. We provide an example OpenBUGS code for each of our models in the technical notes for each chapter. Numerical results presented in this thesis have been generated from R software (version 3.4.4) (R Core Team [2017]) using the package R2OpenBUGS (Thomas [2017]) based on two parallel chains, with each contributing 15 000 MCMC samples and sacrificing the first 5 000 iterations as burn-in. This thesis is closed by a discussion of impact of our research work as well as any limitations of the methodologies, and a brief proposal for future research work.

2

A BAYESIAN DECISION-THEORETIC APPROACH TO USING PRECLINICAL INFORMATION

Summary. Leveraging animal data for a phase I first-in-man trial is appealing yet challenging. A prior based on preclinical information may place large probability mass on values of the dose-toxicity model parameter(s), which appear infeasible in light of data accrued from the ongoing phase I clinical trial. In this paper, we seek to leverage preclinical information to improve decision making in a model-based phase I dose-escalation procedure in oncology. Animal data are incorporated via a robust mixture prior for the parameters of the dose-toxicity relationship. This prior changes dynamically as the trial progresses. After completion of treatment for each cohort, the weight allocated to the informative component based on animal data is updated using a decision-theoretic approach to assess the commensurability of the animal data with the human toxicity data observed thus far. Specifically, we measure commensurability as a function of the utility of optimal predictions, obtained based on animal data alone, for the human responses on each administered dose. The proposed approach is illustrated through several examples and an extensive simulation study. Results show that our proposal can address difficulties in coping with prior-data conflict commencing in sequential trials with a small sample size.

Keywords: Bayesian logistic regression; Decision theory; Phase I dose-finding; Prior-data conflict.

2.1 INTRODUCTION

Phase I oncology trials are performed to characterise the toxicity profile of an anti-cancer therapy in humans. Regulatory authorities require these first-in-man trials to be preceded with preclinical testing of a range of doses involving at least two animal species; moreover, extrapolation of safe doses for a human trial should be based on nonclinical safety studies in the most sensitive and relevant animal species (USFDA [2005]; EMA [2008]). We can therefore reasonably well anticipate some animal data will be made available as preliminary knowledge about toxicity in humans before a phase I clinical trial to be undertaken. Nonetheless, trialists often face the dilemma of using preclinical animal data. On the one hand, there has been a call to design more ethical and efficient phase I clinical trials, basing decision making on all relevant information (O'Quigley et al. [1990]; Neuenschwander et al. [2016a]). On the other hand, using animal data that are inconsistent to the human toxicity could seriously jeopardise the safety of patients to be treated in the phase I clinical trial if failing to disprove the presumption of predictability (Stephen et al. [2007]; Dresser [2009]; Balkwill et al. [2011]). This motivates an investigation of the feasibility of leveraging preclinical animal data into a phase I first-in-man trial.

We see this research question as falling within discussions of the literature on the use of historical data, which could be acquired from external studies under similar circumstances, in a new clinical trial. To date, a number of adaptive methods have been proposed such as power priors (Ibrahim and Chen [2000]; Duan et al. [2006]) and meta-analytic approaches (Neuenschwander et al. [2010]). A primary focus of these methods is on discounting historical data to a proper extent in response to the degree of inconsistency with the newly accrued trial data. In particular, power priors offer a solution of down-weighting external data by raising the historical likelihood to either a fixed or random exponent defined on the interval [0, 1]. Whilst meta-analytic approaches are concerned with between-study heterogeneity that model parameters underpinning the historical trials and that of a new clinical trial are assumed to be conditionally i.i.d. random variables. Historical data are tenuated by large values of the variance that describes parameters which underpin both the historical and new trials. Sophisticated modifications to these methodologies have been proposed. Hobbs et al. [2011, 2012] suggest to explicitly parameterise the commensurability of historical and new data such that a commensurate prior will strongly shrink the new parameter(s) towards the historical parameters when the evidence tends to suggest commensurability. Schmidli et al. [2014] and Neuenschwander et al. [2016a,b]) discuss robust borrowing of information from historical datasets by adopting a mixture prior, which consists of an informative meta-analytic prior and a weakly informative prior, to accommodate scenarios of non-exchangeability of the parameters.

What causes concern in this work is the possibility of erroneous prediction about the human toxicity based upon animal data alone that a dose is safe to administer. Conclusions drawn from preclinical experiments must come along the acknowledged levels of uncertainty. One particular issue encompassing our research question is that preclinical data first need to be translated onto an equivalent human dosing scale. Current practice uses allometric scaling to convert animal doses onto an equivalent human dosing scale through a fixed constant specific to each tested animal species, which is applied to adjust for differences in size (USFDA [2005]). This has incurred controversy, as simple allometry can produce very inaccurate predictions (Sharma and McNeill [2009]). Moreover, despite the best efforts devoted for a very accurate translation across species, conflicts between preclinical and clinical data may still arise due to the intrinsic physiological differences. A plausible solution seems to be formulating a dynamic prior for the human dose-toxicity parameters using the translated animal data, wishing that a flexible down-weighting of animal data, when inconsistent, can be achieved at any stage in an adaptive phase I clinical trial.

In this chapter, we seek to quantify the commensurability of preclinical data with the accumulating human toxicity data using Bayesian decision theory, which has been widely used for clinical trial designs (Brunier and Whitehead [1994]; Muller et al. [2006]; Mandrekar and Sargent [2009]; Saville et al. [2014]). Our context seems to be an ideal setup to apply Bayesian decision methods because investigators are to make a decision, whether or not to incorporate animal data, and the loss has to be set against the risk that more patients may be treated with excessively toxic doses when prior-data conflict commences. We therefore propose a Bayesian decision-theoretic framework to justify adaptive borrowing of preclinical data in an ongoing phase I dose-escalation trial. A set of possible utility functions are specified. Namely, correct prior predictions, made based on animal data alone, will be assigned with a utility value of 1, whilst incorrect prior predictions are penalised with a small utility value. Predictive utility of animal data is then computed across doses of interest after observing patients' outcomes to validate such prior predictions.

The remainder of the chapter is structured as follows. We begin with a motivating example in Section 2.2, and explain how preclinical animal data available on two doses can be represented in a bivariate normal prior for the dose-toxicity parameters of the human trial in Section 2.3. In Section 2.4, we propose a Bayesian decision-theoretic method to adaptively leverage animal data according to a formal assessment of commensurability. We then retrospectively design and analyse the example trial applying the proposed methodology in Section 2.5, and describe a simulation study performed to evaluate the operating characteristics in Section 2.6. Specific focus is to see whether the proposed methodology is responsive to a prior-data conflict in small trials. We close with a discussion of our findings and future research interest in Section 2.7.

2.2 MOTIVATING EXAMPLE

A phase I first-in-man trial of the anticancer therapy AUY922 was performed in 101 patients with the aim of estimating the maximum tolerated dose (MTD) (Sessa et al. [2013]). A set containing nine doses were available for evaluation, $\mathcal{D} = \{$ 2, 4, 8, 16, 22, 28, 40, 54, 70$\}$ mg/m^2. This original dose-escalation trial was conducted with a Bayesian logistic regression model, treating patients sequentially in cohorts of size three (Neuenschwander et al. [2008]). A weakly informative prior (Gelman et al. [2008]) was formulated in light of preclinical data from dog studies, of which the median probabilities of DLTs are about 0.1% and 33% at the doses 2 and 28 mg/m^2, respectively. Fairly limited external information was incorporated such that Bayesian inference will be dominated by the accumulating data from the current trial. We will term such type as "operational priors" hereafter, since they are generally calibrated to ensure that the dose-escalation procedure has acceptable operating characteristics. A dose were to be chosen for the next cohort according to a prespecified probabilistic overdose criterion that

$$d_{\text{sel}}^{(h)} = \max\{d_i : \mathbb{P}(p_i \geqslant 0.33 | x_{\mathcal{H}}^{(h-1)}) \leqslant 0.25\}. \tag{2.2.1}$$

In order to preclude too fast escalation, an additional constraint was imposed that the recommended dose should not exceed a maximum of two-fold increase in the current dose.

This example prompts the following questions: (i) How can we develop a formal approach to incorporate animal data into prior distributions for dose-toxicity model parameters; (ii) How can we dynamically update our prior in response to observed prior-data conflicts, particularly in the early stages of a sequential dose-escalation study when there is much uncertainty. These questions will be taken to motivate the methodology developed in Sections 2.3 and 2.4, to which solutions will be given in Section 2.5 applying the proposed approach.

2.3 REPRESENTING ANIMAL DATA INTO A PRIOR

In this section, we first formally describe a logistic dose-toxicity model that will be considered to guide the Bayesian dose-escalation procedure, and discuss obtaining an informative prior distribution based on available preclinical data on two animal doses.

Let the set $\mathcal{D} = \{d_1, \ldots, d_I; d_{t_1} < d_{t_2} \text{ for } 1 \leqslant t_1 < t_2 \leqslant I\}$ contain all doses available for evaluation. Furthermore, let n_i and r_i denote the number of patients who receive dose $d_i \in \mathcal{D}$ and the number of those who experience a DLT, respectively. We assume that the risk of a DLT on dose d_i, denoted by p_i, increases monotonically with dose and that this relationship can be adequately described using a two-parameter logistic regression model:

$$\begin{aligned} r_i | p_i, n_i &\sim \text{Binomial}(p_i, n_i), \quad \text{for } i = 1, \ldots, I, \\ \text{logit}(p_i) &= \theta_1 + \exp(\theta_2) \log(d_i / d_{\text{Ref}}), \end{aligned} \tag{2.3.1}$$

where d_{Ref} is a predefined reference dose drawn from $\mathcal{D}$. Therefore, θ_1 is the log-odds of toxicity at d_{Ref}. Estimating the model parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$ to perform probabilistic inference about p_i is centre of main interests.

In the Bayesian paradigm, preclinical data, when relevant, can be incorporated into the prior distribution for the model parameters $\boldsymbol{\theta}$, and later updated with the toxicity data from an ongoing phase I human trial. Let $\mathbf{x}_{\mathcal{A}} = \{d_{\mathcal{A}j}; t_j, \nu_j, j = 0, -1, -2, \ldots\}$ denote the animal data that comprise information of binary toxicity outcomes on animal doses $d_{\mathcal{A}j}$ recorded on the original (untranslated) scale: amongst all animal subjects receiving dose $d_{\mathcal{A}j}$, t_j experience a toxicity and ν_j did not. The minimum requirement is that animal data must involve at least two doses and some toxicities must have been observed on the highest dose. We translate the animal doses onto an equivalent human dosing scale by applying, for example, allometric scaling on the basis of body surface area (USFDA [2005]). Thus, we deduce that the risk of a DLT in animals on dose $d_{\mathcal{A}-1}$ is thought to be similar to the risk of a DLT for humans given dose d_{-1}. Similarly, animal dose $d_{\mathcal{A}0}$ is translated to a human equivalent dose of d_0. We note that the pseudo doses d_{-1} and d_0, expressed on the equivalent human dosing scale, are not necessarily contained in the set $\mathcal{D}$. On the basis of the animal data and this interspecies translation, we stipulate independent priors $p_j \sim \text{Beta}(t_j, \nu_j)$ to describe preliminary knowledge about the risks of toxicity at doses d_j (Whitehead and Williamson [1998]). According to these priors, the translated preclinical data on doses $d_{\mathcal{A}j}$ are taken to represent observations on $(t_j + \nu_j)$ patients on dose d_j, $j = -1, 0$.

Given the independent Beta priors on p_{-1} and p_0, we apply the Jacobian transformation to derive the joint probability density of p_i, risk of DLT at a dose $d_i \in \mathcal{D}$, where $i = -1, 0$, and θ_2, which is given as:

$$g_i(p_i, \theta_2) = \frac{1}{p_i(1-p_i)} \exp(\theta_2) \left|\log\left(\frac{d_{-1}}{d_0}\right)\right| \times \prod_{j=-1}^{0} \frac{[1+\exp(-z_{ji})]^{-t_j}[1+\exp(z_{ji})]^{-\nu_j}}{B(t_j, \nu_j)}, \tag{2.3.2}$$

where $z_{ji} = \text{logit}(p_i) + \exp(\theta_2)\log(d_j/d_{\text{Ref}})$, $j = -1, 0$, and $B(a, b)$ in the denominator is the Beta function evaluated at (a, b). We give details on deriving (2.3.2) in the technical notes of this chapter, i.e., Section 2.8.1. The prior distribution of p_i is obtained as:

$$f_i(p_i) = \int_{-\infty}^{+\infty} g_i(p_i, \theta_2) d\theta_2, \tag{2.3.3}$$

and the prior cumulative distribution function for p_i is given as

$$F_i(x) = \int_0^x f_i(p_i) dp_i = \int_0^x \int_{-\infty}^{\infty} g_i(p_i, \theta_2) d\theta_2 dp_i, \quad 0 < x \leqslant 1. \tag{2.3.4}$$

Note that such I prior distributions of p_i can be approximated by a bivariate normal prior for the model parameters θ, by taking the steps stipulated for optimisation as follows. The general idea that one may approximate prior information on the scale of toxicity risks using a prior for dose-toxicity parameters θ was first due to Neuenschwander et al. [2008].

(i) For each dose d_i, the prior for risk of toxicity can be summarised using K percentiles:

$$q_{ik} = \{q(t_{i1}), \ldots, q(t_{iK})\}, \quad i = 1, \ldots, I, \quad k = 1, \ldots, K.$$

In particular, the (100t_{ik})th percentiles of the distribution, denoted by $q(t_{ik})$, can be given by

$$t_{ik} = \mathbb{P}_i(p_i \leqslant q(t_{ik})) = \int_0^{q(t_{ik})} \int_{-\infty}^{\infty} g_i(p_i, \theta_2) d\theta_2 dp_i.$$

Thus with any target $0 < t_{ik} < 1$, we may calculate the percentile as

$$q(t_{ik}) = \inf\{q(t_{ik}) \in (0, 1) : t_{ik} \leqslant \mathbb{P}_i(p_i \leqslant q(t_{ik}))\}.$$

(ii) A bivariate normal prior for θ is found such that the implied percentiles, denoted by q'_{ik}, is in good agreement with q_{ik}. This is an optimisation problem as we aim to minimise the absolute distance between two sets of I × K percentiles, defined as

$$\delta(q_{ik} - q'_{ik}) = \sum_{i,k} |q_{ik} - q'_{ik}|, \quad i = 1, \ldots, I, \quad k = 1, \ldots, K.$$

We note that three (or more) percentiles of distribution for p_i per dose would be needed for a good approximation. In our illustrative examples, we would consider using the median and the limits of 95% credible interval, i.e., $q_{ik} = \{q(0.025), q(0.500), q(0.975)\}$.

We represent the preclinical information on potential dose-toxicity relationship in humans by a bivariate normal prior for θ, which we hereafter refer to as $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$. We would like to add one more note here that the least requirement for leveraging preclinical information is to have animal data on two doses. However, if we had animal data on more than two doses tested in the same preclinical study, we would derive independent beta priors for the probability of toxicity on each tested animal doses, calculate the quantiles of these beta priors, and then approximate them with a bivariate normal prior following the steps developed above. This $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$ can then be used as the prior distribution which may be updated through the Bayes' Theorem, as new human data will accumulate in the phase I trial. However, it is possible that the accumulating human toxicity data may conflict with the prior obtained from animal data.

Figure 2.1 illustrates four possible ways in which preclinical data may conflict with the true (unknown) dose-toxicity relationship in humans. Particularly, preclinical data can constantly (A) over-predict, or (B) under-predict the human DLT risks, or a mixture of (A) and (B) as shown in the panels (C) and (D), across the therapeutic dosing interval where doses have a risk close to the typical target level Γ of 0.25. We want to leverage the preclinical data to inform inferences about θ only if strong evidence shows consistency in toxicity between animals and humans. Another concern is that the consequence of a prior-data conflict in one direction shown in Figure 2.1A may be quite different to that of a conflict in a reversed direction presented in Figure 2.1B. We thus wish to make decisions on using preclinical information informed by the learning about unknown type of prior-data conflict during the course of an adaptive phase I clinical trial.

For a robust borrowing of information, a mixture prior coupling the informative component, $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$ in our case, with a weakly informative component, denoted by

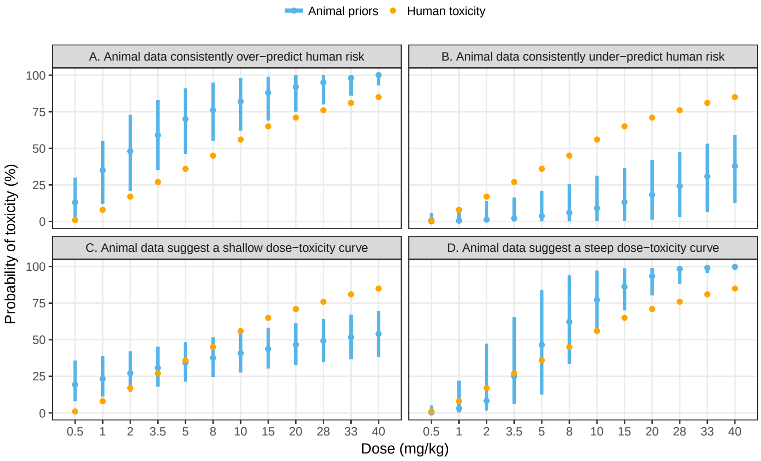

Figure 2.1: Potential commensurability issues for incorporating preclinical information into phase I first-in-man trials. On each dose, the blue point represents the prior median and the endpoints of the bar represent the 95% credible intervals for the risk of DLT, while the orange point suggests the true risk of DLT in humans.

$m_0(\theta)$, is generally adopted to cope with prior-data conflict issues; see Schmidli et al. [2014]; Wadsworth et al. [2018]; Gamalo-Siebers et al. [2017] for example. Denoting the prior weight attributed to the informative component by w, a mixture prior can be written as

$$\mu_0(\theta|\mathbf{x}_{\mathcal{A}}) = w \cdot \pi_0(\theta|\mathbf{x}_{\mathcal{A}}) + (1-w) \cdot m_0(\theta). \tag{2.3.5}$$

Here, a weakly informative $m_0(\theta)$ for routine applied use can be defined to place large probability mass on plausible values of the model parameters. In the motivating example presented in Section 2.2, the original trial was conducted by having $m_0(\theta)$ alone. With a dose-toxicity model in the form of (2.3.1), specification of $m_0(\theta)$ is straightforward:

- set d_{Ref} as a dose chosen from $\mathcal{D}$ that is most likely to be associated with risk of toxicity at the target level $\Gamma \in (0, 1)$;
- calibrate a normal prior for θ_1 so that the prior median equal to the target level Γ and the 95% prior credible interval for the risk of toxicity at dose d_{Ref} is sufficiently wide, say, ranging from 0.01 to 0.95;
- consider a normal prior for θ_2 to accommodate very flat to steep dose-toxicity curves.

In our illustrative examples as well as the simulation study, we will set $\theta_1 \sim N\left(\text{logit}(\Gamma), 2^2\right)$, $\theta_2 \sim N(0, 1^2)$, and a correlation coefficient of 0 for the weakly informative $m_0(\theta)$.

2.4 LEVERAGING PRECLINICAL DATA USING A MIXTURE PRIOR WITH DYNAMICALLY CHOSEN WEIGHTS

In this section, we propose to consider a mixture prior, of which the prior mixture weight for each new patient cohort will be dynamically updated, based on our current knowledge about the commensurability of preclinical data with human toxicity data accured so far. Specifically, developed from (2.3.5), we consider for cohort h of the trial that

$$\mu_0^{(h)}(\theta|\mathbf{x}_{\mathcal{A}}) = w^{(h)} \cdot \pi_0(\theta|\mathbf{x}_{\mathcal{A}}) + (1 - w^{(h)}) \cdot m_0(\theta), \tag{2.4.1}$$

where $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$ and $m_0(\theta)$ are as defined, and the formulations do not depend on the cohort number h. The prior mixture weights will be dynamically assessed for the flexibility of discarding preclinical data entirely when they are found to be completely inconsistent with human toxicity at any stage of the trial. The cohort-specific prior mixture weight $w^{(h)} \in [0, 1]$ attributed to $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$ determines the amount of preclinical animal data to be leveraged. This fraction can be interpreted as investigators' confidence about the commensurability of animal data with the human toxicity. In the following, we develop a Bayesian decision-theoretic framework for choosing $w^{(h)}$ during an ongoing trial. The objective is to optimise learning about unknown prior-data conflict without undermining the safety of patients.

2.4.1 *Assessment of commensurability using a Bayesian decision theoretic approach*

Define the random variable Y_i as the binary DLT outcome of a new patient assigned dose $d_i \in \mathcal{D}$, such that $Y_i = 1$ when a patient experiences a DLT, and $Y_i = 0$ otherwise. Furthermore, let $\tilde{y}_i$ denote the realisation of Y_i. The prior predictive probability mass function (Vehtari and Ojanen [2012, 2014]) of Y_i given the animal data is written as:

$$\begin{aligned} \mathbb{P}_i(Y_i = \tilde{y}|\mathbf{x}_{\mathcal{A}}) &= \int f(\tilde{y}|p_i) f_i(p_i|\mathbf{x}_{\mathcal{A}}) dp_i \\ &= \int_0^1 p_i^{\tilde{y}} \cdot (1 - p_i)^{1-\tilde{y}} f_i(p_i|\mathbf{x}_{\mathcal{A}}) dp_i, \qquad \text{for } \tilde{y} \in \{0, 1\}. \end{aligned} \tag{2.4.2}$$

Before a patient's response has been observed, one could use the prior predictive distribution in (2.4.2) to derive a prediction for Y_i, labelled η_i, where $\eta_i \in \{0, 1\}$. Then, after the patient has been treated and followed-up, we can compare the prediction

Table 2.1: Cross-tabulation of utilities for the predicted versus actual human binary DLT outcomes.

		Observation ($\tilde{y}$)	
		No-DLT	DLT
Prior prediction ($\hat{\eta}$)	No-DLT	u_{00}	u_{10}
	DLT	u_{01}	u_{11}

with the response actually observed. Table 2.1 lists the possible configurations of the predicted and observed outcomes. Let $U(\tilde{y}_i, \eta_i)$ denote the utility function that can take values of $u_{\ell s}$ contained in the cells of Table 2.1. Before Y_i is observed, the optimal prior prediction based on the animal data alone is

$$\hat{\eta}_i = \arg \max_{\eta_i \in \{0,1\}} \sum_{\tilde{y}_i \in \{0,1\}} U(\tilde{y}_i, \eta_i) \mathbb{P}_i(Y_i = \tilde{y}_i). \tag{2.4.3}$$

Fouskakis and Draper [2002] suggest a metric quantifying discrepancy between predicted and actual values needs to be defined for assessing the accuracy of a model's prediction. In our problem, we believe the commensurability is closely linked with predictability of human toxicity using animal data alone. We would therefore wish to establish a formal assessment of predictive accuracy of preclinical data, and interpret this as a quantification method of commensurability between toxicities of the drug to animals and humans. As the observed human toxicity outcomes accrue along with time, the assessment is best to be carried out dynamically. To be more specifically, once a patient has been treated and their response observed, the optimal prior predictions made based upon preclinical data will be compared with the actual observations accrued thus far to assess the predictive accuracy of animal data. In the following, we will define a dynamic metric for commensurability in the context of an adaptive phase I clinical trial.

Let a subset $\mathcal{D}'(h-1) \subseteq \mathcal{D}$ contain the doses that have been tested after cohort $(h-1)$ has been enrolled and treated. At each administered dose $d_i \in \mathcal{D}'(h-1)$, we summarise the counts of patients by the first $(h-1)$ cohorts, denoted by $n_{\ell s}^{(h-1)}$, for whom the preclinical data led to an optimal prediction $\hat{\eta} = \ell$ and their observed outcome was $\tilde{y} = s$, with $\ell \in \{0, 1\}$ and $s \in \{0, 1\}$. For $i = 1, \ldots, I$, define the predictive utility of preclinical information on an administered dose d_i as

$$G_i^{(h-1)}(u_{\ell s}, n_{\ell s}^{(h-1)}) = \sum_{\ell=0}^{1} \sum_{s=0}^{1} u_{\ell s} n_{\ell s}^{(h-1)}, \tag{2.4.4}$$

where the utilities $u_{\ell s}$ remain constant across doses and cohorts and are derived as $U(\tilde{y} = \ell, \hat{\eta}_i = s) = u_{\ell s}$. Values of $u_{\ell s}$ are stipulated to penalise incorrect predictions and reward correct ones. In what follows, we therefore allocate a utility of 0 (1) to an incorrect (correct) prediction, that is, in our notation, we set $u_{10} = 0$ and $u_{00} = u_{11} = 1$. For an incorrect prediction of DLT, we consider a utility of $u_{01} = b$, $0 < b < 1$ since predicting a no DLT as DLT, unlike predicting a DLT as no DLT, will not undermine patients' safety but only bring in additional caution. We do not allocate a utility as large as 1 in light of the potential drawback that a conservative dose-escalation might be resulted in.

By the time, commensurability of the given preclinical information with human toxicity on a particular dose can be characterised as

$$c_i^{(h-1)}(u_{\ell s}, n_{\ell s}^{(h-1)}) = \frac{G_i^{(h-1)}(u_{\ell s}, n_{\ell s}^{(h-1)})}{\sum_{l=0}^{1} \sum_{s=0}^{1} n_{\ell s}^{(h-1)}}, \tag{2.4.5}$$

where the denominator is the maximum utility that would be achieved if all prior predictions were correct. A measurement describing the predictive accuracy of the animal data for the human DLT data at doses of most interest from the first $(h-1)$ cohorts can be defined by taking the average of $c_i^{(h-1)}(u_{\ell s}, n_{\ell s}^{(h-1)})$ across $T^{(h-1)}$ doses in $\mathcal{D}'(h-1)$ that have been administered so far and those either falling in the neighbourhood of, or more toxic than, the current best estimate of MTD:

$$\kappa^{(h-1)}(u_{\ell s}, n_{\ell s}^{(h-1)}) = \frac{1}{T^{(h-1)}} \sum_i c_i^{(h-1)}(u_{\ell s}, n_{\ell s}^{(h-1)}). \tag{2.4.6}$$

The interpretation is that we would be more interested in commensurability of prior predictions, or say, animal data, and human toxicity locally at the target dose. The so-called neighbourhood throughout will be confined as one dose level removed from an estimated target dose in our implementation. Considering the assessment at high doses in $\mathcal{D}$ is helpful for distinguishing penalisation to animal data that have either over-predicted or under-predicted human toxicity. Here, we particularly prefer not to compute the predictive accuracy at low doses when they are estimated to be much safer the target dose. This is because, in such a scenario, differences between the animal and human dose-toxicity relationships are present but too small to result in discrepancies between prior predictions based on animal data and human outcomes. Including those very safe low doses in $\mathcal{D}'(h-1)$ will then lead to an undesirably large value of the average predictive accuracy. The quantity $\kappa^{(h-1)}$ will then be used to determine the cohort-specific mixture weight $w^{(h)}$ to be attributed to the informative animal prior, $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$, such that value of preclinical information and knowledge about the commensurability can be exploited.

2.4.2 *Choosing an appropriate tuning parameter*

One concern arising here is that $\kappa^{(h-1)}(u_{\ell s}, n_{\ell s}^{(h-1)})$ assessed using the accumulating human data can be a noisy estimator. This issue could perhaps be outstanding at early stages of a phase I trial when few doses have been tried and few patients have been tested. With accrual of the human data, the assessment on predictive accuracy becomes more convincing. Taking account of this, we wish to come up with a sensible formulation of the mixture weight $w^{(h)}$ in (2.4.1) defined as a function of $\kappa^{(h-1)}$. We restrict our attention to a power-law functional relationship

$$w^{(h)} = \{\kappa^{(h-1)}(u_{\ell s}, n_{\ell s}^{(h-1)})\}^{\lambda}, \tag{2.4.7}$$

in which λ is a time-varying tuning parameter, which increases from 0 to 1 as more human data accumulate. We consider two *ad-hoc* choices for λ in the following.

A simple proposal is to relate λ explicitly to the trial information time N/n_h, for example, in a convex decreasing function on the interval of [1, N] such as $\lambda = \sqrt{N/n_h}$, where N and n_h are the maximum sample size of the trial and the total number of patients in the first h cohorts, respectively. This power law function for λ has been used in a different context by Thall and Weathen for response adaptive randomisation schemes (Thall and Wathen [2007]). We note this is not uncommon to alter a scientific assessment so as to meet an ethical preference.

As an alternative, one may wish to define λ in a way to suggest how noisy our prediction of $c_i^{(h)}$ could be, looking ahead to the next patient cohort. More specifically, we suppose $d^{(h)}$ is the dose selected for cohort h based on certain criterion. Optimal prior prediction for the corresponding toxicity outcome $\hat{\eta}_i$ of a patient receiving a known $d^{(h)}$ is unambiguous, while the observable toxicity outcome is random that Y_i (here, $i = d^{(h)}$) takes value of either 0 or 1. The predictive accuracy $c_i^{(h)}$ is hence a discrete random variable: if $\hat{\eta}_i = 0$, predictive utility can take the values u_{00} or u_{10}, with probabilities $\Pr(Y_i = 0)$ and $\Pr(Y_i = 1)$, respectively. Similarly, if $\hat{\eta}_i = 1$, predictive utility takes values u_{01} or u_{11}, with probabilities $\Pr(Y_i = 0)$ and $\Pr(Y_i = 1)$, respectively. These probabilities will be approximated by the current best estimate of the risk of toxicity at dose $d^{(h)}$, say, the prior predictive probability when $h = 1$ and the posterior probability when $h \geqslant 2$. We can then derive standard deviation of the predictive accuracy given randomness of Y_i to be observed from the newly treated patient. Let H be the maximum number of cohorts that trialists have planned. We will then simulate binary outcomes for patients of cohort $(h+1)$ to H, assuming that they would all be treated with the current best estimate of MTD which is dose $d^{(h)}$. Finally, we stipulate

$$\lambda = \frac{\sigma\{c_i^{(h)}(u_{\ell s}, \tilde{n}_{\ell s}^{(h)})\}}{\sigma\{c_i^{(H)}(u_{\ell s}, \tilde{n}_{\ell s}^{(H)})\}}, \tag{2.4.8}$$

where $\tilde{n}_{\ell s}^{(\cdot)}$ denote the counts of patients actually treated on the trial and the future patients whose toxicity outcomes are simulated. This formulation of λ reflects how noisy our prediction of $c_i^{(h)}$ is, compared to what would be the case at the end of the trial when data from N patients have been collected. It takes account of trial information time implicitly in that λ is given a large value at the beginning but it converges to 1 by the end of the phase I trial.

2.5 DESIGN AND ANALYSIS FOR THE EXAMPLE TRIAL INCORPORATING ANIMAL DATA

In this section, we illustrate how the proposed method can incorporate preclinical information into a phase I dose-escalation study. Here, we define the target dose as one associated with a risk of toxicity of 0.25.

2.5.1 *Prior distributions based on preclinical information*

For illustrative purpose, we suppose that preclinical animal data are available on doses 0.1 and 2.7 mg/kg which have been used in 30 dogs each, and that there are 1 and 17 toxicities observed from each dosage group. Following the allometric scaling approach that standardises body weight (BW) by body surface area (BSA), one may calculate the equivalent human doses of these two animal doses using

$$\text{Equivalent human dose (mg/m}^2\text{)} = \text{Animal dose (mg/kg)} \times \left(\frac{\text{BW}}{\text{BSA}}\right)_{\text{Animal}},$$

as specified in the USFDA [2005] draft guideline *Estimating the Maximum Safe Starting Dose in Initial Clinical Trials for Therapeutics in Adult Healthy Volunteer*. Appropriate for dogs, we substitute the BW and BSA with fixed constants 10 and 0.5, respectively, and obtain the equivalent human doses as 2 mg/m^2 (labelled dose d_{-1}) and 54 mg/m^2 (labelled dose d_0), respectively. We further express the dogs data as pseudo-observations such that $p_{-1} \sim \text{Beta}(1, 29)$ and $p_0 \sim \text{Beta}(17, 13)$.

Since we do not directly have data on $d_i \in \mathcal{D}$, we derive the prior distributions for p_i, based on the dogs data, under an assumption that the dose-toxicity model follows a logistic regression in log dose. Following the steps outlined in Section 2, the 2.5th, 50th and 97.5th percentiles of the marginal prior distributions for p_i are presented in Figure 2.2A, together with the corresponding fitted probabilities from the approximated bivariate normal prior for θ that

$$\theta|x_{\mathcal{A}} \sim \text{BVN}\left(\begin{pmatrix} -0.524 \\ 0.147 \end{pmatrix}, \begin{pmatrix} 0.151 & -0.008 \\ -0.008 & 0.001 \end{pmatrix}\right).$$

As we can see from Figure 2.2A – 2.2B, our hypothetical preclinical information suggests that dose 16 and 22 mg/m^2 have a risk of toxicity in humans close to the target level of 0.25, when no robustification is concerned. We further summarise each of the prior distributions for the risks of toxicity in Figure 2.2B, with three interval probabilities: (i) probability of underdosing, where $p_j \in [0, 0.16)$, (ii) probability that p_j lies in the target interval $[0.16, 0.33)$, and (iii) probability of overdosing, where p_j exceeds 0.33 (Neuenschwander et al. [2008]). Figure 2.2C presents the prior probability density curves for the low doses. In our illustrative examples, we set dose 4 mg/m^2 as the starting dose, as it appears to be safe given animal data with $x_{\mathcal{A}}$ that $\mathbb{P}(p_2 < 0.1|x_{\mathcal{A}}) = 0.825$.

To evaluate the effective sample size (ESS) (Morita et al. [2008]) of each marginal prior distributions for each p_i on each dose implied by $\pi_0(\theta|x_{\mathcal{A}})$, we approximate each prior with a $\text{Beta}(a, b)$ distribution, where the parameters are chosen to match the first two moments of the prior. The ESS is then found as $(a + b)$. Table 2.2 lists the prior ESSs suggesting the preclinical data provide information on the risks of toxicity on low doses equivalent to that which would be obtained from 16.4 – 20.1 humans on the risk of toxicity on medium to high doses. This may overwhelm accumulating data from an ongoing phase I trial with small sample size. It is thus important to see whether the proposed method can effectively down-weight preclinical information in situations when it is inconsistent with observed human toxicity data. We will now illustrate the behaviour of dose-escalation procedures using our method to integrate preclinical data via several data examples.

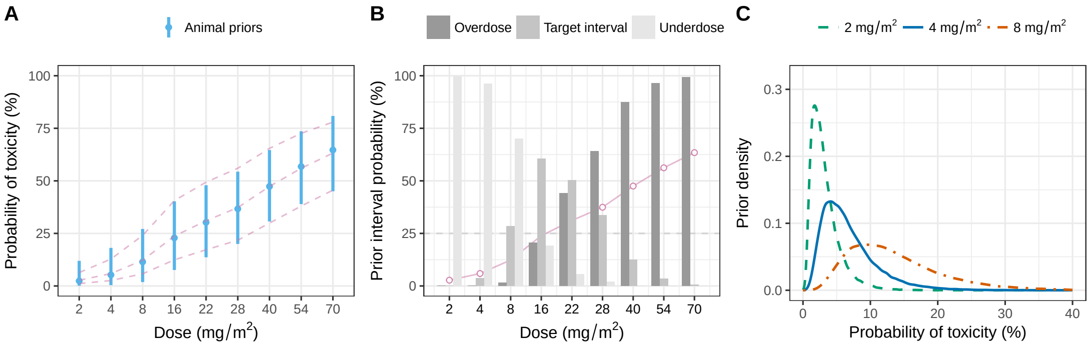

Figure 2.2: Summaries of priors based on preclinical information. Panel A shows median and 95% CI of the marginal prior distributions for the probability of toxicity in blue bars, together with the fitted probabilities in pink dashed lines from the bivariate normal prior $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$ found with our optimiser. Panel B gives an overview about interval probabilities, where the background red curve indicates the prior medians for probability of toxicity per dose. Panel C presents prior densities for the risks of toxicity at candidate starting doses.

Table 2.2: Effective sample sizes of marginal prior distributions for risk of toxicity based on animal data summarised by $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$.

	Dose (mg/m^2)								
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
	2	4	8	16	22	28	40	54	70
Prior means	0.034	0.065	0.120	0.211	0.267	0.314	0.393	0.463	0.526
Prior std dev.	0.022	0.038	0.061	0.089	0.101	0.109	0.117	0.119	0.118
ESS	67.1	41.7	27.6	20.1	18.2	17.2	16.5	16.4	16.9
a	2.3	2.7	3.3	4.2	4.9	5.4	6.5	7.6	8.9
b	64.8	39.0	24.3	15.9	13.3	11.8	10.0	8.8	8.0

2.5.2 *Hypothetical data examples*

With Figure 2.2B, we see that doses up to 16 mg/m^2 comply with the escalation criterion defined in (2.2.1) if the preclinical information is to be fully incorporated. We will now adopt a mixture prior with dynamically chosen weight to implement the Bayesian dose-escalation procedure. A utility of $u_{01} = 0.6$ is specified for quick discounting of preclinical information that predicts a no-DLT as DLT. Based on the animal data summarised by the bivariate normal prior $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$ defined in equation (2.4.1) and the utilities $u_{00} = u_{11} = 1, u_{10} = 0, u_{01} = 0.6$, the optimal predictions for DLT outcomes in humans are: patients receiving doses d_i in the set $\{2, 4, 8, 16\}$ will not experience a DLT ($Y_i = 0$); patients who receive doses $\{22, 28, 40, 54, 70\}$ will experience a DLT.

We simulated three data examples consistent with (i) negligible prior-data conflict, (ii) a prior-data conflict, where preclinical information under-estimates the DLT risk in humans, and (iii) a prior-data conflict, where preclinical information over-estimates the DLT risk in humans. In particular, each of these data examples are one of the realisations setting doses 22, 4 and 40 mg/m^2 as the true MTD, respectively. Said in another way, the risk of toxicity at these doses across data examples is set to be 25%.

Figure 2.3 presents the trial history and dynamic updates of the prior mixture weight attributed to $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$ for these three data examples, assuming that patients were recruited in cohorts of size three. Interim dose recommendations were made using the proposed robust Bayesian model stipulating the tuning parameter in the form of (2.4.8). We assume each hypothetical trial recruits a maximum of 11 cohorts; that is, the maximum sample size $N = 33$. After completion of treatment for each patient cohort, the cohort-specific mixture weights will be updated with the newly accrued data via the Bayes' Theorem as the posterior probability of relevance, denoted by $w_*^{(h)}$,

$$\mu^{(h)}(\theta|\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{H}}^{(h)}) = w_*^{(h)} \cdot \pi^{(h)}(\theta|\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{H}}^{(h)}) + (1 - w_*^{(h)}) \cdot m^{(h)}(\theta|\mathbf{x}_{\mathcal{H}}^{(h)}), \tag{2.5.1}$$

where $\mathbf{x}_{\mathcal{H}}^{(h)}$ denote the human data collected from the first h cohorts.

Across the three hypothetical trials, we see that the given preclinical animal data are incorporated for interim dose recommendations with a prior mixture weight of $w^{(h)} = 1$ until the first discrepancy occurs between the prior predictions, made based on animal data together with the investigators' utilities, and the observed human outcomes. This occurs in the forth cohorts in data examples 1 and 3, and the first

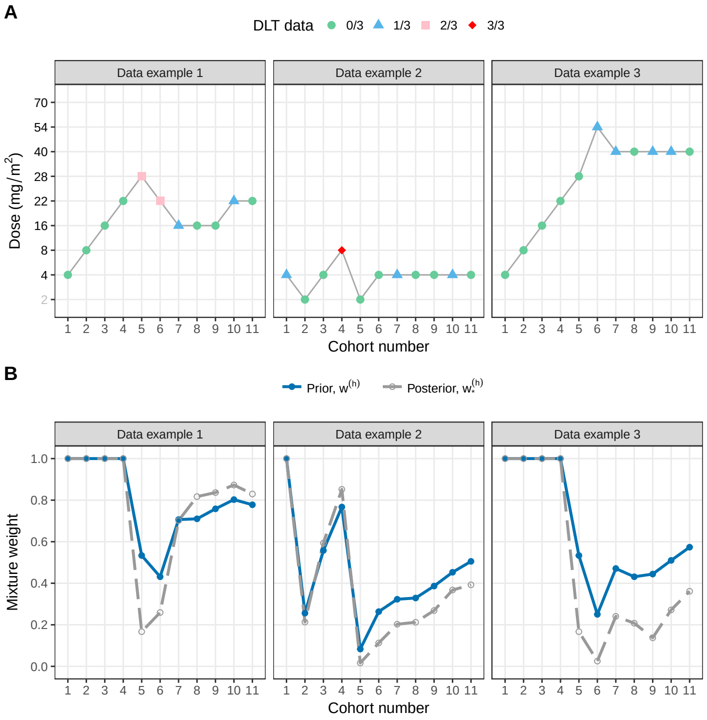

Figure 2.3: Trajectory of dose recommendations (Panel A) and dynamic update of mixture weight attributed to preclinical information (Panel B) during the course of each hypothetical data example.

cohort in data example 2. Indeed, a disagreement between predictions and observed data can take a few cohorts to emerge, even when there is a conflict between the human dose-toxicity curve and our opinions (illustrated in Figure 2.2A) based on animal data alone. This is because low doses in $\mathcal{D}$ will generally be very safe. Thus, there is little chance of observing a DLT in the first few cohorts treated with low doses, meaning prior predictions of no-DLT outcomes based on animal data will be correct.

The prior mixture weight decreases after the first disagreement is observed, which is generally followed with a small trend of increase for the last several cohorts. This is because the predictive accuracy $\kappa^{(h-1)}(u_{\ell s}, n_{\ell s}^{(h-1)})$ assess at the target dose and its neighbourhood is unlikely to vary substantially while the tuning parameter λ reduces to 1 as the trial progresses. In data examples 1–3, the phase I trial terminated declaring the dose 22, 4, and 40 mg/m^2 as the target dose, respectively. Moreover, for the last cohort of each scenario, we see the prior mixture weight is assessed to be the largest that $w^{(11)} = 0.778$ in data example 1 for prior-data consistency, and smaller that $w^{(11)} = 0.505$ and 0.574 in data examples 2 and 3, respectively. We will now be particularly focused onto data examples 2 and 3 to evaluate property of the proposed method in situations of a prior-data conflict.

In data example 2, the increase in $w^{(h)}$ to 0.767 for cohort $h = 4$ after the rapid discounting of preclinical information at the beginning of the trial was due to correct predictions on the two lowest doses. Preclinical information was penalised drastically after erroneously predicting DLT as no-DLT for all three patients treated in the forth cohort, that the prior probability of relevance was quantified as 0.247 for the fifth cohort. Data example 3 shows how the procedure reacts to a data-conflict when the DLT risks in humans are much lower than predicted by the animal data. Cohorts 1–5 escalate up to 28 mg/m^2 and no DLTs are observed, which turned out to contradict the prior predictions based on animal data on doses 22 and 28 mg/m^2. Consequently, $w^{(h)}$ drops from 1 to 0.294 for cohort $h = 6$. Reduced borrowing from animal data resulted in an escalation to dose 54 mg/m^2, at which one out of three patients was observed with a DLT. A de-escalation to 40 mg/m^2 then took place, and estimate of the target dose eventually converges at this dose level. Finally, we note using our approach leads to compromises between full pooling and complete discarding of preclinical animal data. Results of these assessments are available in Figure 2.6 in the Supplementary Materials.

2.5.3 *Specifying a run-in period*

As was illustrated in Section 2.5.2, the proposed approach described in Section 2.4 tends to implement full borrowing of preclinical information in early stages of the phase I trial when human toxicity outcomes at low doses can be correctly predicted. However, it is counterintuitive that we assign full weight to preclinical data in the early stages of a trial when few human data are available to assess commensurability. This is particularly true since we know the agreement between prior predictions, based on animal data alone, and the observed human data is an artefact of starting the trial with very safe doses rather than a reflection of a genuine agreement. If the dose-toxicity relationship in humans has a very steep slope, placing full weight on the animal prior could lead to overly aggressive escalation and observation of unexpected DLTs once we enter the therapeutic dosing range.

To this end, we consider implementing a constrained version of the proposed decision-theoretic approach by introducing a run-in period. We note this is an *ad-hoc* solution that we will set the prior mixture weight $w^{(h)} = 0$ until the first discrepancy between prior predictions and actually observed outcomes. In other words, a weakly informative operational prior $m_0(\theta)$, instead of an informative animal prior $\pi_0(\theta|x_{\mathcal{A}})$, will be used to start off the phase I dose-escalation trial. Preclinical information will then come into play, serving as a component of a mixture prior $\mu_0^{(h)}(\theta)$ rather than guiding the escalation scheme on its own. The ethical stance here, which is clearly associated with efficiency terms, is that, until the data tend to suggest the presence of difference between toxicity of the drug in animals and humans, scepticism towards preclinical information holds. On another note, when trials may be designed with additional early stopping rules to conclude on a MTD, robust inference is achieved in that available animal data are unlikely to override the sparse data accumulated from a phase I human trial at any stage.

To reach a fair comparison about properties of the constrained and unconstrained versions of our methodology, we present three new data examples in Figure 2.4 that have been simulated from same parameter settings used for those in Figure 2.3: a vector of binary outcomes on each dose were simulated and then sampled without replacement as each new patient was assigned to a dose in the dose-escalation study. For instance, in terms of data examples with the same label, the first patient assigned a dose of 28 mg/m^2 in Figure 2.3 will have the same simulated toxicity outcome as the first patient assigned to the same dose in Figure 2.4. Therefore, we can observe what may be resulted in by having a run-in period to constrain the proposed approach.

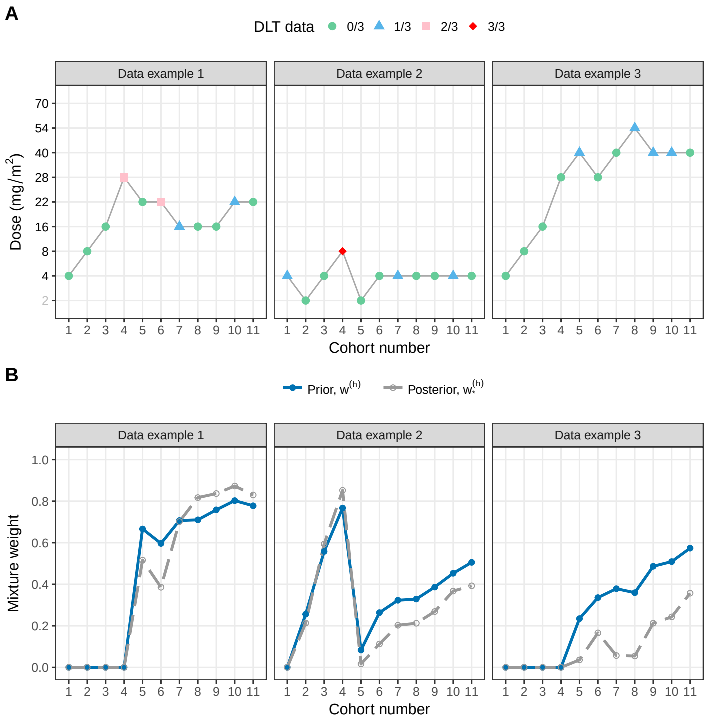

Figure 2.4: Trajectory of dose recommendations (Panel A) and dynamic update of mixture weight attributed to preclinical information (Panel B) during the course of each hypothetical data example with a two-stage design, with a run-in period characterised in the first stage and dose-escalation procedure driven by a mixture prior in the second stage.

As shown in data examples 1 and 3 presented in Figure 2.4, this permits skipping the dose 22 mg/m^2 so long as the probabilistic overdose control criterion is met after observations on dose 16 mg/m^2. In data example 3, allocation of doses to patients entering the trial changed, along with decreased borrowing of preclinical information at the interims. No particular changes are noticed for data example 2, representative to trials in which prior prediction for outcome(s) of the first cohort is not entirely correct: a prior mixture weight $w^{(h)} < 1$ was allocated to the preclinical component for every following cohort where an interim dose recommendation is needed. By the end of each simulated trial implementing the constrained version of our approach, we observed the prior mixture weights $w^{(11)} <$ have been assessed to be the same value as those implementing the unconstrained version. This is because of our way to simulate the human toxicity data presented in Figures 2.3 and 2.4, together with the assessment, which let data to speak, is solely based on the compatibility of preclinical and clinical trial data. Nevertheless, we noticed that dose escalations tend to be less restrictive especially at the early stages of the simulated trials, leading to different interim dose recommendations.

2.6 SIMULATION STUDY

2.6.1 *Basic settings*

In this section, we evaluate the operating characteristics of phase I dose-escalation trials which are designed and conducted basing inferences on the robustified mixture prior described in Sections 2.3 and 2.4 with the decision-theoretic weights (with and without a run-in period). Comparisons are made with trials basing inferences on mixture priors with fixed weights. We stipulate the following alternative dose-escalation procedures for comparison:

- Procedure A: no run-in period; Bayesian mixture prior with decision-theoretic weights
- Procedure B: run-in period; Bayesian mixture prior with decision-theoretic weights
- Procedure C: no run-in period; Bayesian mixture prior with fixed weights $w^{(h)} = 0.5$, for $h = 1, 2, \ldots$
- Procedure D: Bayesian informative prior with fixed weights $w^{(h)} = 1$, for $h = 1, 2, \ldots$

- Procedure E: Bayesian model not permitting borrowing of preclinical data $w^{(h)} = 0$, for $h = 1, 2, \ldots$

Here we note that the justification of prior mixture weight set as 0.5 in Procedure C was somewhat subjective but presumptively to be sufficient to respond to a prior-data conflict in most cases.

Further, we assume the preclinical information described in Section 2.5.1 is available prior to the phase I dose-escalation trial. To evaluate the behaviour of fast discounting in cases of prior-data inconsistency, we simulate phase I trials with very small sample size: there are in total seven cohorts planned (each comprising three patients) for evaluating doses contained in set $\mathcal{D}$. For reasons previously stated, dose 4 mg/m^2 is chosen as the starting dose for patients in the first cohort of the trial. Interim dose recommendations are made according to criterion (2.2.1), with the same caveats on maximum two-fold escalation step defined. Trials end when all 21 patients have been treated and observed, or after any cohort h, if the lowest dose is found to be excessively toxic that $\mathbb{P}(p_1 \geqslant 0.33 | \mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{H}}^{(h)}) > 0.25$. These two subsets of simulated trials will later be referred to as *complete* or *stopped early* trials, respectively.

Phase I oncology trials are simulated under the eight human toxicity scenarios shown in Table 3.4. In Scenario 3, the true probabilities of toxicity are consistent with the prior median estimates obtained from the animal data, summarised in Figure 2.2A and derived assuming the dose-toxicity curve follows a logistic regression model. In none of the other toxicity scenarios were human DLT probabilities derived from a logistic regression model. For each Bayesian procedure A–E, results will be presented based on 1000 simulated trials per toxicity scenario.

At the end of a *complete* trial, we estimate the MTD with the point estimate (say, posterior median) of the DLT risk, denoted by $\tilde{p}_i$, defining that

$$\hat{d}_M = \arg \min_{d_i \in \mathcal{D}_c} |\tilde{p}_i - 0.25|,$$

where $\mathcal{D}_c \subseteq \mathcal{D}$ comprises all the doses that have been administered to humans during the trial and comply with our overdose control criterion. For each scenario, we report the percentage of *complete* trials (which make a definitive declaration of a dose as MTD), and the percentage of trials *stopped early* for safety without a MTD declaration. For the group of *complete* trials, we report the PCS of procedures A–E. We use the optimal nonparametric design as a benchmark for comparisons. Moreover, we report the average number of patients allocated to each dose across the 1000 simulated trials.

Table 2.3: Simulation scenarios for the true probability of DLT in humans. The figure in bold indicates the dose closest to the true MTD.

	Dose (mg/m^2)								
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
	2	4	8	16	22	28	40	54	70
Scenario 1	0.11	**0.25**	0.35	0.41	0.47	0.52	0.58	0.63	0.70
Scenario 2	0.08	0.16	**0.25**	0.35	0.42	0.45	0.53	0.60	0.70
Scenario 3	0.02	0.05	0.14	**0.25**	0.35	0.42	0.51	0.60	0.68
Scenario 4	0.03	0.05	0.10	0.16	**0.25**	0.32	0.40	0.48	0.55
Scenario 5	0.001	0.005	0.03	0.10	0.16	**0.25**	0.38	0.50	0.60
Scenario 6	0.01	0.02	0.05	0.08	0.11	0.14	**0.25**	0.37	0.47
Scenario 7	**0.35**	0.42	0.60	0.75	0.82	0.88	0.91	0.94	0.97
Scenario 8	0.001	0.005	0.01	0.02	0.04	0.05	0.10	0.16	**0.25**

2.6.2 *Results*

Numerical results are recorded in Table 2.4 in the Supplementary Materials (Section 2.9), where properties of our approach considering a small utility of $u_{01} = 0.2$ have also been evaluated. Figure 2.5 visualises the operating characteristics of simulated phase I dose-escalation trials (using $u_{01} = 0.6$), in which interim inferences were made based on Procedures A – E, respectively. We can see that the constrained version of our methodology, Procedure B, performs reasonably well across all the toxicity scenarios. Procedure A performs similarly in nearly all scenarios except Scenario 8, where the true MTD is the highest dose. Animal data in this scenario over-predict the toxicity in humans; incorporating them thus results in a more conservative dose-escalation scheme. Accordingly, in Scenario 8, Procedure A (with no run-in period) was more cautious and treated more patients in the early stages at quite safe lower doses, at which the prior predictions were mostly correct; while in contrast, Procedure B (with a run-in period) allowed quicker escalation than Procedure A, since the given animal data come into play only after escalation has reached to high doses and enough human data have been cumulated for assessment of commensurability. This explains the results in Panel (ii) that on average only 1.3 out of 21 patients were allocated to 70 mg/m^2 with Procedure A, while slightly more (2.4 out of 21) were to this true MTD when using Procedure B. Substantial difference in the PCS, displayed in Panel (i), between these two procedures in Scenario 8 is therefore not surprising. We observe this difference is even larger if taking a smaller utility such as $u_{01} = 0.2$. Same explanation will be given to address the different behaviours of Procedures A and B in Scenario 6. In the other scenarios, similar operating characteristics were observed, as overall commensurability assessed by the end of a trial simulated from the same toxicity scenario converges.

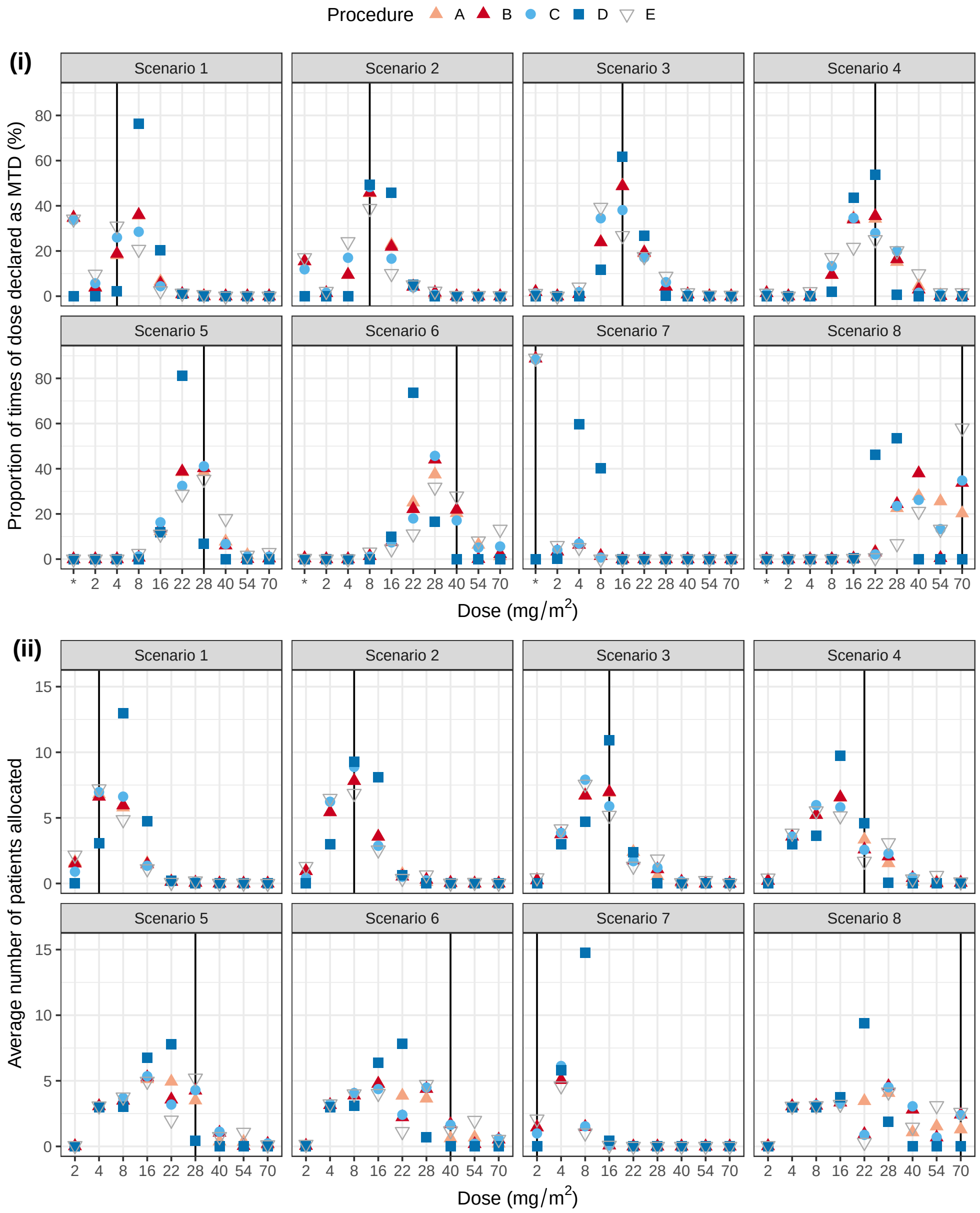

Figure 2.5: Operating characteristics of phase I clinical trials designed using the dose-escalation Procedures A - E. The vertical black line indicates the true MTD in humans in each simulation scenario.

Comparisons between Procedures B and C illustrate one advantage of our decision-theoretic approach to setting the prior mixture weight: Procedure B permits increased borrowing of information in cases of prior-data consistency, while behaving equally well in response to a prior-data conflict. For example, an increase in PCS is seen from 38.1% to 48.7% in Scenario 3. However, there is no free lunch. Our proposal, especially the unconstrained version without stipulating a run-in period, experiences problem in discounting quickly the animal data that provide correct prior predictions but suggest a different dose to be the MTD in humans. Precisely, this happens when differences between dose-toxicity relationships are present but too small to result in discrepancies between predictions based on animal data and human outcomes, and sometimes also due to choice of the user utilities for prediction. This is illustrated by Scenario 8, where we read that Procedures A claimed dose 40 mg/m^2 to be the MTD for most of the simulated trials. Based on additional simulations, of which the results are not shown here, Procedure B appears to be more advantageous than Procedure C with a fixed prior weight chosen to be smaller than 0.5 in scenarios of a severe prior-data conflict.

Procedure D outperforms all other Procedures for using preclinical data only in Scenarios 2 – 4, because of the prior-data consistency to a certain degree. Whereas, in Scenarios 1 as well as 5 – 8, a dose-escalation procedure based on Procedure D appears to be very restrictive since the inconsistent animal data dominated inferences. Procedure B compromises between incorporating and discarding entirely the animal data is demonstrated to be flexible enough across all these toxicity scenarios: it comes close to Procedure D in the prior-data consistency scenarios, while shows similar property to Procedure E in the prior-data conflict scenarios.

Readers may notice there is a high proportion of trials *stopped early* in Scenario 1 when implementing the dose-escalation designs except Procedure E: about 35% of the simulated trials were terminated before reaching the maximum sample size planned. This is because we apply a very constrained criterion for early stopping, $\mathbb{P}(p_1 \geqslant 0.33 | \mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{H}}^{(h)}) > 0.25$. It means that the first-in-man trial has to be terminated, if potentially more than 25% of the patients receiving the lowest dose, 2 mg/m^2 which is not the starting dose in our numerical examples, will have more than 33% of possibility to experience a DLT. When trialists expect fewer trials will be stopped earlier to correctly select the starting dose, 4 mg/m^2, as the MTD, it will work to either (i) relax the definition of overdose, increasing 0.33 in the criterion to a larger value such as 0.45, or (ii) increase the acceptable bound, for example, setting $\mathbb{P}(p_1 \geqslant 0.33 | \mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{H}}^{(h)}) > 0.5$. Alternatively, investigators may set a very safe dose, lower than

2 mg/m^2 in our numerical examples, below the starting dose. Once DLTs may be observed from patients receiving the starting dose, it will de-escalate to the back-up dose, giving chances to escalate from this dose to the true MTD at later stages of the trial. For allocation of patients to doses available for evaluation and average sample size used, our approach (Procedures A and B) permits more patients to be treated with the true MTD in scenarios of prior-data consistency, while it performs equally well with Procedures C and E in scenarios of a prior-data conflict.

Results shown in Figure 2.5 were derived with the tuning parameter specified in the form of (2.4.8) for Procedures A and B. However, we can confirm the validity of its alternative setting the tuning parameter in relation with trial information time explicitly that $\lambda = \sqrt{N/n_h}$, which produces consistent numerical results and will not modify the conclusions. We have also run simulations setting the maximum trial sample size as 33 (i.e., 11 patient cohorts) and 45 (i.e., 15 patients cohorts). We present the simulation results with a larger sample size than discussed above in Figure 2.8, as we can see similar conclusions are drawn, with the exception that differences between Procedures A and B for the PCS in scenario 8 becomes smaller with increase of the sample size. This is because as the number of human cohorts increases, we eventually escalate up to reach the highest dose in Scenario 8 when using Procedure A. On the other hand, if we had more informative preclinical animal data than the one currently in use for illustration purpose, adopting Procedure B becomes more advantageous than Procedure C for its assessment of commensurability to determine the amount of information to borrow.

In addition to the eight toxicity scenarios, we have also evaluated the performance of our dose-escalation procedures in situations, where the true dose-toxicity curve in humans is very steep or shallow. We do not present the numerical results from these evaluations, as they are well expected. In the investigated scenario with a steep human dose-toxicity curve, say, there is a sudden increase in the DLT risk from one dose level to the next level, animal data underpredict the human DLT risk at the high doses and will therefore be discounted very quickly by using our methodology. With the escalation criterion set based on the interval probability of overdose, it is unlikely that the dose will be escalated to the level that has very high DLT risk. In contrast, in the investigated scenario with a shallow human dose-toxicity curve, say, the difference between DLT risks at two doses that are next to each other is about 5%, the shape of the dose-toxicity relationship learned from the animal data will offer information to discern differences between the doses. When there is a clear discrepancy between the true DLT risks and the predicted DLT risks at the human doses, animal data will also be down-weighted to certain extent, depending on the magnitude of such discrepancy as well as the type of erroneous prior prediction.

2.7 DISCUSSION

The question of using historical data in a new trial has been discussed elsewhere, but in the context of leveraging preclinical information in a phase I first-in-man trial, there are unique circumstances to be taken care of. Indeed, the challenge is to address potential prior-data conflicts, arising from the intrinsic difference between toxicity of a drug to animals and humans, which emerge in a sequential trial planned with a small sample size. Particularly, “small” is meant in relation to the prior effective sample size. In this chapter, we have outlined solutions for translating preclinical animal data recorded on their original scale to predict and facilitate efficiently estimating the toxicity in humans, especially at the doses available to be administered to patients, and proposed a Bayesian decision-theoretic approach to using such information in an ongoing phase I dose-escalation trial in an adaptive way. Commensurability of the translated animal data with the newly accrued human toxicity data is successively assessed to determine a sensible amount of borrowing. A formal quantification of the commensurability was proposed with respect to how correct the prior predictions based on animal data may be, by comparing them with the observed outcomes to be collected later on. Incorrect predictions will be penalised by giving a small utility value to quickly discount animal data during the course of a phase I clinical trial.

In current practice of phase I first-in-man trials, a strategy called sentinel dosing is often considered so that one human subject in the first cohort is dosed in advance of the full study. This ensures fewer human subjects to be impacted in situations when DLTs would manifest very quickly. This would fit the use of our approach nicely: the outcome observed from this sentinel subject may be compared with the prior prediction, obtained using animal data, to learn about the commensurability between toxicity in animals and humans, and therefore decide how much prior weight would be allocated to the animal data. When a run-in period would be incorporated to use our approach, having a sentinel subject will not influence directly in recommendation of a suitable dose for the next patient(s). The benefit is simply to allow sufficient time between dosing for a more ethical first-in-man trial.

Illustrative examples and the simulation study have shown that the proposed methodology leads to sensible borrowing of preclinical information to aid decision making in a phase I clinical trial, and is responsive to a prior-data conflict emerges any time during the trial. We note that obtaining robust inference does not seem to be readily possible in a most basic kind of borrowing based on the Bayes' Theorem, which incorporates preclinical information as what it is entirely. Conventionally, if desired to be used in a new phase I clinical trial, preclinical animal data would be down-weighted to contain least amount of information in the first place, so as to avoid overriding data from the trial. Whereas, our approach provides a possibility to borrow strength from preclinical animal data adaptively. It is a developed version of mixture prior with feasible mid-course modification of the prior mixture weight. As we have observed from the simulations comparing our approach with its origin, potential benefit includes the increased borrowing in cases of prior-data consistency and the capability of discounting any inconsistent prior even quicker.

When formulating the research problem, we have assumed that animal data were available from two interesting doses, as quite a few preclinical animal studies are conducted to evaluate the toxicity on a qualitative basis. However, this should not be taken as a restriction of applying the proposed methodology. When richer preclinical animal data are available from a number of preclinical *in vivo* studies performed in one species, information may be synthesised using meta-analysis to derive the prior predictive distribution for probability of toxicity per dose to be evaluated in humans, and used to make optimal prior predictions for assessing the commensurability of the synthesised animal data with human toxicity data. The discussion of using animal data collected from preclinical studies involving multiple animal species is beyond the scope of this paper. In such a more challenging case, we may wish to allocate larger prior mixture weights to animal species that are more relevant to humans than others, whereas the decision-theoretic approach proposed at present does not allow us to draw the distinction. This is where we look toward for the future work to extend the methodology. We are also currently pursuing the use of animal pharmacokinetic information by establishing a Bayesian dose-exposure-toxicity model in light of the growing interest in better understanding and characterisation of the dose-toxicity relationship based on mechanisms of pharmacological action (USFDA [2003]; José and Stephen [2009]).

2.8 TECHNICAL NOTES

2.8.1 *Deriving the marginal probability density function for* p_j

We consider to express the preclinical animal data for predicting the risks of toxicity at human doses as pseudo-data. Thus, at these two pseudo dose levels $j = -1, 0$, uncertainty surrounding the risks of toxicity p_j could be described using Beta distributions with parameters t_j and ν_j. The joint prior probability density function (pdf) of p_{-1} and p_0 is given by

$$f(p_{-1}, p_0) = \prod_{j=-1}^{0} \frac{p_j^{t_j-1}(1-p_j)^{\nu_j-1}}{B(t_j, \nu_j)},$$

where $B(\cdot,\cdot)$ is the beta function.

Given the logistic dose-toxicity model, the joint pdf $f(p_{-1}, p_0)$ can be expressed in terms of the model parameters θ_1 and θ_2 via Jacobian transformation,

$$h(\theta_1, \theta_2) = f(p_{-1}, p_0) \times \frac{\partial(p_{-1}, p_0)}{\partial(\theta_1, \theta_2)}. \tag{2.8.1}$$

From

$$\log\left(\frac{p_j}{1-p_j}\right) = \theta_1 + \exp(\theta_2)\log(d_j/d_{\text{Ref}}),$$

we can easily derive

$$\frac{\partial p_j}{\partial \theta_1} = p_j(1-p_j) \quad \text{and} \quad \frac{\partial p_j}{\partial \theta_2} = p_j(1-p_j)\exp(\theta_2)\log(d_j/d_{\text{Ref}}).$$

Thus, the joint prior pdf of θ_1 and θ_2 can be written as

$$h(\theta_1, \theta_2) = \exp(\theta_2)\left|\log\left(\frac{d_{-1}}{d_0}\right)\right| \times \prod_{j=-1}^{0} \frac{p_j^{t_j}(1-p_j)^{\nu_j}}{B(t_j, \nu_j)},$$

where the two pseudo doses d_{-1} and d_0 correspond to the lowest and highest human doses in our context. Substituting the p_j with the logistic model parameters, we can write this joint prior pdf more explicitly:

$$h(\theta_1, \theta_2) = \exp(\theta_2)\left|\log\left(\frac{d_{-1}}{d_0}\right)\right| \times \prod_{j=-1}^{0} \frac{[1+\exp(-z_j)]^{-t_j}[1+\exp(z_j)]^{-\nu_j}}{B(t_j, \nu_j)},$$

where $z_j = \theta_1 + \exp(\theta_2)\log(d_j/d_{\mathrm{Ref}})$.

By applying Jacobian transformation again, we can further derive the joint prior pdf of p_i and θ_2; for $i = 1, \ldots, I$,

$$g_i(p_i, \theta_2) = h(\theta_1, \theta_2) \times \frac{\partial(\theta_1, \theta_2)}{\partial(p_i, \theta_2)}.$$

With

$$\begin{cases} \theta_1 = \log\left(\dfrac{p_i}{1-p_i}\right) - \exp(\theta_2)\log(d_i/d_{\mathrm{Ref}}) \\ \theta_2 = \theta_2 \end{cases}$$

we can write

$$\begin{aligned} \frac{\partial\theta_1}{\partial p_i} = \frac{1}{p_i(1-p_i)}, &\qquad \frac{\partial\theta_1}{\partial\theta_2} = 0, \\ \frac{\partial\theta_2}{\partial p_i} = 0, &\qquad \frac{\partial\theta_2}{\partial\theta_2} = 1, \end{aligned}$$

such that

$$\begin{aligned} g_i(p_i, \theta_2) &= h(\theta_1, \theta_2) \times \frac{\partial(\theta_1, \theta_2)}{\partial(p_i, \theta_2)}, \\ &= \frac{1}{p_i(1-p_i)} \cdot \exp(\theta_2)\left|\log\left(\frac{d_{-1}}{d_0}\right)\right| \times \prod_{j=-1}^{0} \frac{[1+\exp(-z_{ji})]^{-t_j}[1+\exp(z_{ji})]^{-\nu_j}}{B(t_j, \nu_j)}, \end{aligned}$$

where $z_{ji} = \theta_1 + \exp(\theta_2)\log(d_j/d_{\mathrm{Ref}})$.

Because θ_1 in z_j can be expressed with θ_2 and p_i given the logistic regression model that

$$z_{ji} = \log\left(\frac{p_i}{1-p_i}\right) + \exp(\theta_2)\log\left(\frac{d_j}{d_i}\right),$$

the joint prior pdf $g_i(p_i, \theta_2)$ can therefore be parameterised with only p_i and θ_2. The marginal probability density function for p_j, the risk of toxicity at dose d_i, $i = 1, \ldots, I$, can then be derived by integrating out the nuisance parameter θ_2:

$$f_i(p_i) = \int g_i(p_i, \theta_2) d\theta_2.$$

2.8.2 *Implied percentiles on the scale of* p_j*, given a bivariate normal prior for* θ

For $\log\left(\frac{p}{1-p}\right) = z$, the 95% credible interval for p is bounded by $\left(\frac{\exp(z_L)}{1+\exp(z_L)}, \frac{\exp(z_U)}{1+\exp(z_U)}\right)$ should we have known the lower and upper limits of z. Here z can be seen as a transformed random variable, as following our parameterisation $z = \theta_1 + \exp(\theta_2)\log(d/d_{\text{Ref}})$.

The first two moments of the transformed random variable z

The expectation for z is $\mathbb{E}(z) = \mathbb{E}[\theta_1 + \exp(\theta_2)\log(d/d_{\text{Ref}})] = \mathbb{E}(\theta_1) + \log(d/d_{\text{Ref}})\mathbb{E}(\exp(\theta_2))$. By Taylor expansion, we know

$$\begin{aligned}\mathbb{E}(\exp(\theta_2)) &\approx \exp(\mathbb{E}(\theta_2)) + \frac{1}{2}\exp(\mathbb{E}(\theta_2)) \cdot \text{Var}(\theta_2) \\ &= \exp(\mathbb{E}(\theta_2))[1 + \frac{1}{2}\text{Var}(\theta_2)] \\ &\approx \exp\left[\mathbb{E}(\theta_2) + \frac{1}{2}\text{Var}(\theta_2)\right].\end{aligned}$$

The last step follows the Taylor approximation $\exp(x) \approx 1 + x$, which works well for small x. Having $x = \frac{1}{2}\text{Var}(\theta_2)$ leads to $\exp(\mathbb{E}(\theta_2))[1 + x] \approx \exp(\mathbb{E}(\theta_2) + x)$. Thus, the first moment for z is approximated as

$$\mathbb{E}(z) = \mathbb{E}(\theta_1) + \log(d/d_{\text{Ref}})\exp\left[\mathbb{E}(\theta_2) + \frac{1}{2}\text{Var}(\theta_2)\right]. \qquad (2.8.2)$$

Since $z^2 = \theta_1^2 + 2\theta_1\exp(\theta_2)\log(d/d_{\text{Ref}}) + \exp(2\theta_2)[\log(d/d_{\text{Ref}})]^2$, the second moment is then given by

$$\begin{aligned}\mathbb{E}(z^2) &= \mathbb{E}(\theta_1^2) + 2\log(d/d_{\text{Ref}}) \cdot \mathbb{E}(\theta_1 \cdot \exp(\theta_2)) + [\log(d/d_{\text{Ref}})]^2 \cdot \mathbb{E}(\exp(2\theta_2)) \\ &= \text{Var}(\theta_1) + [\mathbb{E}(\theta_1)]^2 + 2\log(d/d_{\text{Ref}})[\text{Cov}(\theta_1, \exp(\theta_2)) + \mathbb{E}(\theta_1)\mathbb{E}(\exp(\theta_2))] \\ &\qquad + [\log(d/d_{\text{Ref}})]^2 \cdot \mathbb{E}(\exp(2\theta_2)),\end{aligned} \qquad (2.8.3)$$

while we have

$$[\mathbb{E}(z)]^2 = [\mathbb{E}(\theta_1)]^2 + 2\log(d/d_{\text{Ref}}) \cdot \mathbb{E}(\theta_1)\mathbb{E}(\exp(\theta_2)) + [\log(d/d_{\text{Ref}})]^2[\mathbb{E}(\exp(\theta_2))]^2.$$

Thus,

$$\begin{aligned}\mathrm{Var}(z) &= \mathbb{E}(z^2) - [\mathbb{E}(z)]^2\\ &= \mathrm{Var}(\theta_1) + 2\log(d/d_{\mathrm{Ref}})\cdot\mathrm{Cov}(\theta_1, \exp(\theta_2))\\ &\qquad + [\log(d/d_{\mathrm{Ref}})]^2[\mathbb{E}(\exp(2\theta_2)) - \mathbb{E}(\exp(\theta_2))]^2\\ &= \mathrm{Var}(\theta_1) + 2\log(d/d_{\mathrm{Ref}})\cdot\mathrm{Cov}(\theta_1, \exp(\theta_2)) + [\log(d/d_{\mathrm{Ref}})]^2\cdot\mathrm{Var}(\exp(\theta_2)).\end{aligned} \tag{2.8.4}$$

For the $\mathrm{Cov}(\theta_1, \exp(\theta_2))$ in (2.8.4), with Stein's Lemma, it holds that

$$\begin{aligned}\mathrm{Cov}(\theta_1, \exp(\theta_2)) &= \mathbb{E}(\exp(\theta_2))\cdot\mathrm{Cov}(\theta_1, \theta_2)\\ &\approx \exp\left[\mathbb{E}(\theta_2) + \frac{1}{2}\mathrm{Var}(\theta_2)\right]\cdot\mathrm{Cov}(\theta_1, \theta_2).\end{aligned}$$

For the $\mathrm{Var}(\exp(\theta_2))$ in (2.8.4),

$$\begin{aligned}\mathrm{Var}(\exp(\theta_2)) &= \mathbb{E}(\exp(2\theta_2)) - [\mathbb{E}(\exp(\theta_2))]^2\\ &\approx \exp(2\mathbb{E}(\theta_2) + 2\mathrm{Var}(\theta_2)) - \exp(2\mathbb{E}(\theta_2) + \mathrm{Var}(\theta_2))\\ &= \exp(2\mathbb{E}(\theta_2) + \mathrm{Var}(\theta_2))\cdot\exp(\mathrm{Var}(\theta_2)) - \exp(2\mathbb{E}(\theta_2) + \mathrm{Var}(\theta_2))\\ &= \exp(2\mathbb{E}(\theta_2) + \mathrm{Var}(\theta_2))[\exp(\mathrm{Var}(\theta_2)) - 1].\end{aligned}$$

The lower and upper limits of z

With (2.8.2) and (2.8.4), the lower and upper limits of z are

$$\begin{aligned}z_L &= \mathbb{E}(z) - 1.96\sqrt{\mathrm{Var}(z)},\\ z_U &= \mathbb{E}(z) + 1.96\sqrt{\mathrm{Var}(z)}.\end{aligned}$$

Obtaining the implied percentiles denoted by q'_{jk}, we can then easily code up the optimiser used to find a bivariate normal prior $\pi(\boldsymbol{\theta})$ given the prior probabilities q_{jk} obtained following steps described in Section 2.

2.8.3 *OpenBUGS code for implementation*

```
model{
# sampling model
```

```
for(j in 1:Ncohorts){
lin[j] <- theta[1] + exp(theta[2])∗log(doseH[j]/dRef)
logit(pTox[j]) <- lin[j]
NtoxH[j] ~ dbin(pTox[j], NsubH[j])
}

for(i in 1:MdoseH){
lin.star[i] <- theta[1] + exp(theta[2])∗log(doseH[i]/dRef)
logit(pTox.star[i]) <- lin.star[i]

pCat[i, 1] <- step(pTox.cut[1] - pTox.star[i])
pCat[i, 2] <- step(pTox.cut[2] - pTox.star[i])
                  - step(pTox.cut[1] - pTox.star[i])
pCat[i, 3] <- step(1 - pTox.star[i]) - step(pTox.cut[2] - pTox.star[i])
}

theta[1:2] ~ dmnorm(thetaMu[which, 1:2], thetaPrec[which, 1:2, 1:2])
which ~ dcat(wMix[1:2])
# to monitor the exchangeability probability
# in the course of the new human trial
for(k in 1:2){
prob.ex[k] <- equals(which, k)
}


thetaMu[1, 1:2] ~ dmnorm(PriorA[1:2], thetaPrec[1, 1:2, 1:2])
cov.A[1, 1] <- PriorA[3]
cov.A[1, 2] <- PriorA[4]
cov.A[2, 1] <- cov.A[1, 2]
cov.A[2, 2] <- PriorA[5]
thetaPrec[1, 1:2, 1:2] <- inverse(cov.A[1:2, 1:2])


thetaMu[2, 1:2] ~ dmnorm(Prior.mw[1:2], thetaPrec[2, 1:2, 1:2])
cov.rb[1, 1] <- pow(Prior.sw[1], 2)
cov.rb[2, 2] <- pow(Prior.sw[2], 2)
cov.rb[1, 2] <- Prior.sw[1]*Prior.sw[2]*Prior.corr
cov.rb[2, 1] <- cov.rb[1, 2]
thetaPrec[2, 1:2, 1:2] <- inverse(cov.rb[1:2, 1:2])

}
```

2.9 SUPPLEMENTARY MATERIALS

2.9.1 *Data examples for no borrowing or full pooling of animal information*

In Section 2.5.2, we simulated three hypothetical phase I clinical trials to exemplify interim dose recommendations, using the proposed method to leverage preclinical data without undermining patients' safety. Here, we show in Figure 2.6 how doses would have been recommended in an alternative Bayesian dose-escalation procedure driven by either an operational prior $m_0(\theta)$ or an animal prior $\pi_0(\theta|\mathbf{x}_{\mathcal{A}})$.

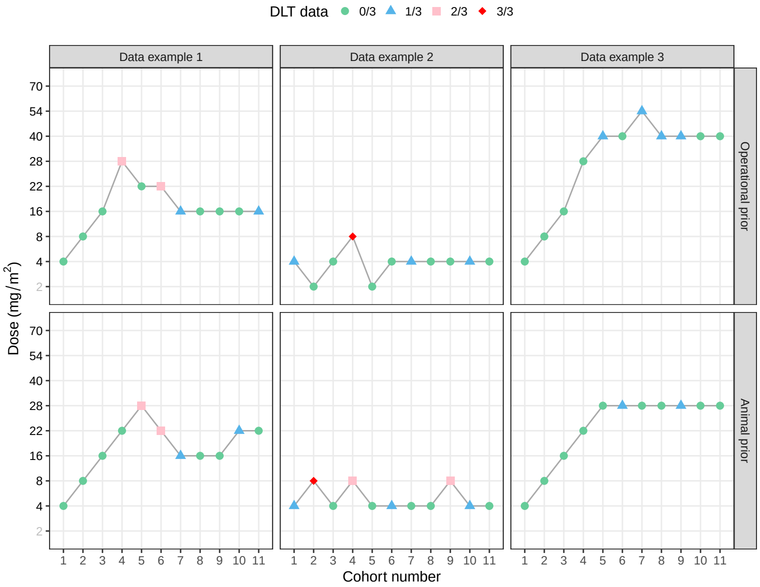

Figure 2.6: Trajectory of dose recommendations under alternative Bayesian dose-escalation procedures.

These new data examples presented in Figure 2.6 were simulated from the same parameter settings used for those presented in Figures 2.3 and 2.4 of the main manuscript for a fair comparison: a vector of binary outcomes on each dose were simulated and then sampled without replacement as each new patient was assigned to a dose in the dose-escalation study. For example, looking at data examples with the same label, the first patient receiving dose 28 mg/m^2 in Figures 2.3 and 2.4 and Figure 2.6, involved in different Bayesian dose-escalation procedures will have the same simulated binary toxicity outcome. We may therefore observe what could have been the consequence by adopting different priors in a Bayesian dose-escalation procedure.

From this comparison, we can see the impact of leveraging the preclinical data on decision making in an adaptive phase I clinical trial. Using our approach leads to compromises between full pooling and complete discard of preclinical animal data. In the simulated trials labelled with data example 1, we observe that behaviours of the trial using our approach is similar with that implemented with Bayesian approach fully incorporating animal data in the prior, as under this prior-data consistency scenario a large prior mixture weight will be attributed to animal data based upon our assessment of commensurability. Advantages of using our approach are also evident in scenarios of a prior-data conflict: unlike a trial with animal data fully incorporated, less patients were allocated with overly toxic dose 8 mg/m^2 in data example 2, while more patients will have chance to escalate to a true target dose which is dose 40 mg/m^2 in data example 3.

2.9.2 *Numerical results of all evaluate scenarios*

The performance of trials using BLRM-guided dose-escalation Procedures A – E are compared with that of the optimal non-parametric benchmark design by Maccario et al. [2002]. The optimal design is defined using the 'complete' toxicity profile of each patient, created by assuming there are $J_{i^\star}$ clones of a patient given doses spanning the dosing set $\mathcal{D}_{i^\star}$. A toxicity tolerance thereshold ϵ_n is generated from $U[0,1]$ for the nth patient, which determines the corresponding toxicity outcome at the jth dose as

$$R_{jn} = \mathbb{1}(\epsilon_n \leqslant p_{i^\star j}), \quad 1 \leqslant n \leqslant N, \quad 1 \leqslant j \leqslant J_{i^\star},$$

where $\mathbb{1}(\cdot)$ is the indicator function. An unbiased estimate for $p_{i^\star j}$ is thus $\bar{R}_j(N) = \frac{1}{N}\sum_{n=1}^{N} R_{jn}$ for a trial of which the maximum sample size is N. Consequently, the estimated MTD under the benchmark design is

$$\hat{d}_M^{\text{opt}} = \arg \min_{j=1,\ldots,J_{i^\star}} |\bar{R}_j(N) - 0.25|.$$

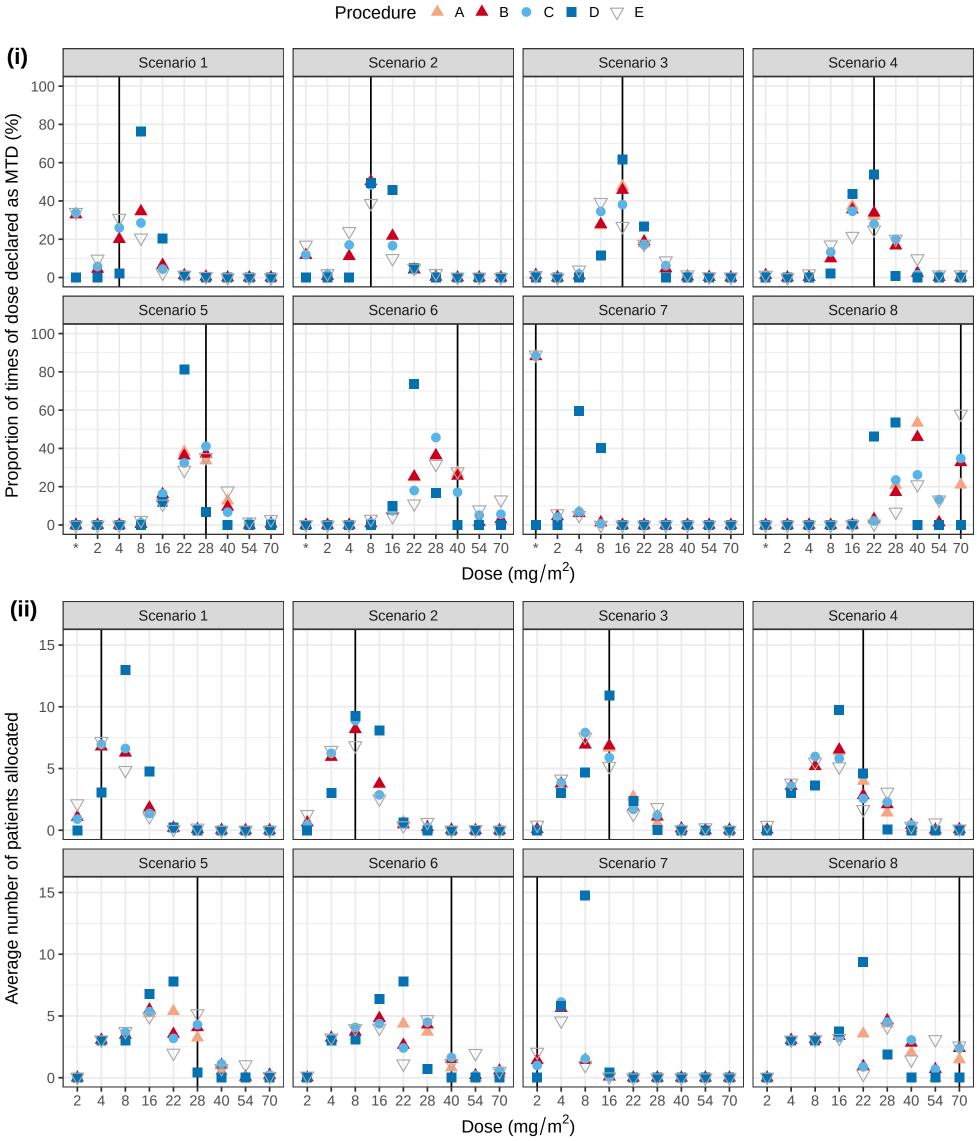

Figure 2.7: Operating characteristics of phase I clinical trials designed using the dose-escalation Procedures A - E, where the tuning parameter is stipulated explicitly relating to the trial information time for Procedures A and B. The vertical black line indicates the true MTD in humans in each simulation scenario.

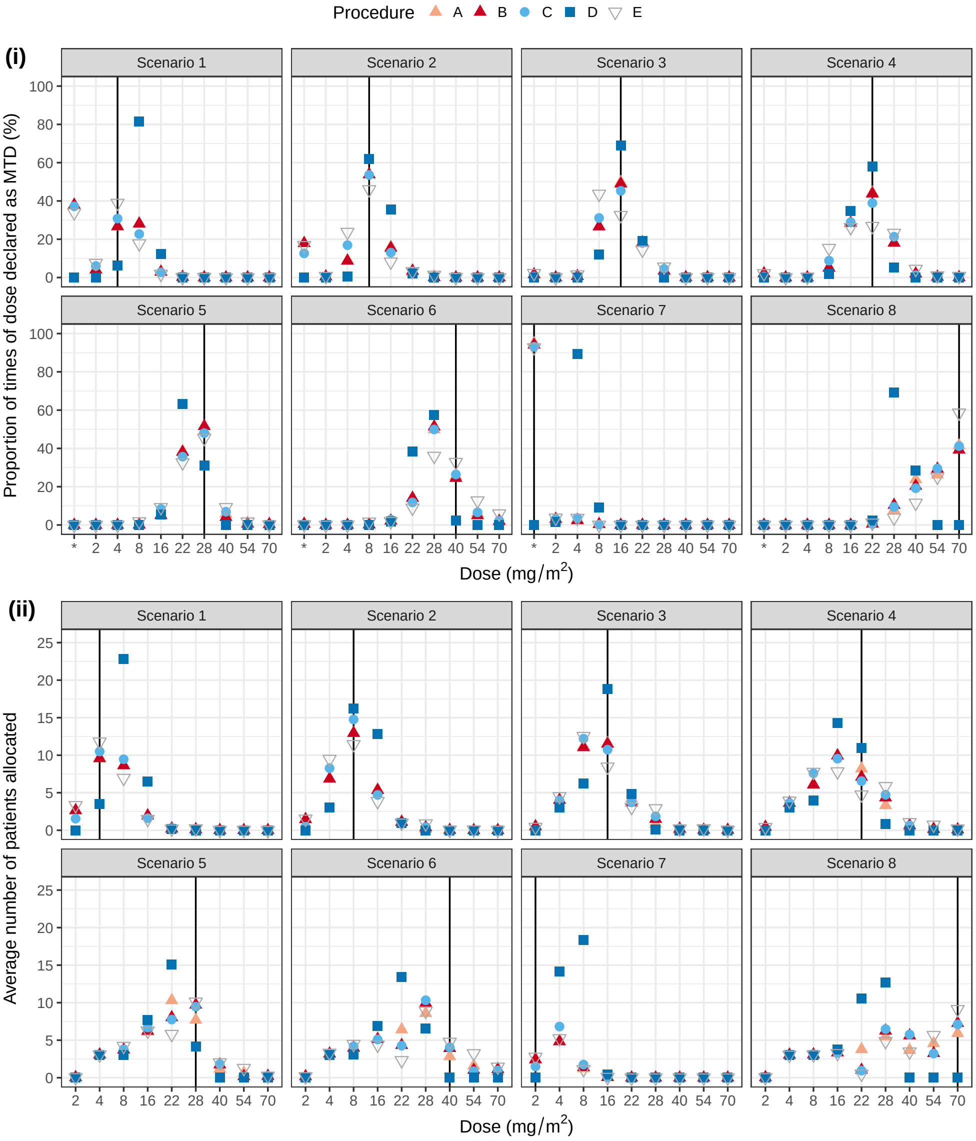

Figure 2.8: Operating characteristics of phase I clinical trials, setting the maximum trial sample size as 33 (i.e., 11 patient cohorts), designed using the dose-escalation Procedures A - E. The vertical black line indicates the true MTD in humans in each simulation scenario.

Table 2.4: Comparison of alternative analysis models in terms of the percentage of selecting a dose as MTD at the end of the trials, percentage of early stopping for safety, average patient allocation, and average number of patients with toxicity. For each simulated trial, we specify the maximum sample size as 21.

Sc.	Design			% dose declared as MTD & average patient allocation										DLT	$\bar{N}$
				$d_{i^\star 1}$ 2	$d_{i^\star 2}$ 4	$d_{i^\star 3}$ 8	$d_{i^\star 4}$ 16	$d_{i^\star 5}$ 22	$d_{i^\star 6}$ 28	$d_{i^\star 7}$ 40	$d_{i^\star 8}$ 54	$d_{i^\star 9}$ 70	None		
1			pTox	0.11	**0.25**	0.35	0.41	0.47	0.52	0.58	0.63	0.70			
	Optimal		Sel	18.2	**54.2**	19.8	5.6	1.5	0.4	0.1	0	0			
	A	$u_{01} = 0.6$	Sel	4.2	**18.2**	35.7	6.6	0.7	0.1	0	0	0	34.5		
			Pts	1.6	6.8	5.8	1.6	0.2	0	0	0	0		4.7	16.0
		$u_{01} = 0.2$	Sel	4.2	**18.2**	35.8	6.4	0.8	0.1	0	0	0	34.5		
			Pts	1.6	6.8	5.8	1.5	0.2	0	0	0	0		4.7	15.9
	B	$u_{01} = 0.6$	Sel	3.8	**18.8**	36.0	5.6	1.0	0	0	0	0	34.8		
			Pts	1.5	6.6	6.0	1.5	0.1	0.1	0	0	0		4.6	15.8
		$u_{01} = 0.2$	Sel	3.8	**18.8**	35.9	5.7	1.0	0	0	0	0	34.8		
			Pts	1.5	6.6	6.0	1.4	0.2	0.1	0	0	0		4.7	15.8
	C		Sel	5.7	**26.0**	28.5	4.4	1.3	0.4	0	0	0	33.7		
			Pts	0.9	7.0	6.6	1.4	0.1	0.1	0	0	0		4.8	16.1
	D		Sel	0	**2.2**	76.4	20.3	1.1	0	0	0	0	0		
			Pts	0	3.1	13.0	4.7	0.2	0	0	0	0		7.3	21.0
	E		Sel	9.7	**31.0**	20.7	2.4	1.3	0.7	0	0.1	0	34.1		
			Pts	2.2	7.2	4.9	1.1	0.1	0.2	0	0	0		4.3	15.7
2			pTox	0.08	0.16	**0.25**	0.35	0.41	0.45	0.52	0.58	0.63			
	Optimal		Sel	4.3	28.1	**39.8**	20.3	4.9	1.7	0.7	0.1	0.1			
	A	$u_{01} = 0.6$	Sel	1.4	9.5	**45.6**	22.7	4.4	0.8	0.1	0	0	15.5		
			Pts	1.0	5.4	7.8	3.6	0.7	0.1	0	0	0		4.5	18.6
		$u_{01} = 0.2$	Sel	1.4	9.5	**46.3**	21.8	4.6	0.9	0	0	0	15.5		
			Pts	1.0	5.4	7.8	3.6	0.7	0.1	0	0	0		4.5	18.6
	B	$u_{01} = 0.6$	Sel	1.4	9.5	**45.9**	21.9	4.2	1.6	0	0	0	15.5		
			Pts	1.0	5.4	7.8	3.6	0.5	0.3	0.1	0	0		4.5	18.7
		$u_{01} = 0.2$	Sel	1.4	9.5	**46.0**	21.4	5.2	0.9	0.1	0	0	15.5		
			Pts	1.0	5.4	7.8	3.6	0.6	0.3	0.1	0	0		4.5	18.8
	C		Sel	1.5	17.0	**48.3**	16.6	3.9	0.7	0.1	0	0	11.9		
			Pts	0.4	6.2	8.9	2.9	0.5	0.2	0	0	0		4.6	19.1
	D		Sel	0	0	**49.3**	45.7	5.0	0	0	0	0	0		
			Pts	0	3.0	9.3	8.1	0.6	0	0	0	0		5.8	21.0
	E		Sel	2.1	24.1	**38.8**	10.0	5.0	2.2	0.3	0.3	0.2	17.0		
			Pts	1.3	6.5	6.9	2.5	0.4	0.6	0	0.1	0		4.2	18.3
3			pTox	0.02	0.05	0.14	**0.25**	0.35	0.42	0.51	0.60	0.68			
	Optimal		Sel	0	1.0	24.5	**46.3**	21.5	5.3	1.3	0.1	0			
	A	$u_{01} = 0.6$	Sel	0	1.0	24.0	**49.0**	19.0	4.1	0.9	0.1	0	1.9		
			Pts	0.2	3.8	6.7	6.9	2.4	0.6	0	0	0		3.9	20.6
		$u_{01} = 0.2$	Sel	0	1.0	24.9	**47.1**	21.2	3.9	0	0	0	1.9		
			Pts	0.2	3.8	6.8	7.0	2.5	0.4	0	0	0		4.5	20.7
	B	$u_{01} = 0.6$	Sel	0	1.0	23.9	**48.7**	19.4	4.3	0.8	0	0	1.9		
			Pts	0.2	3.8	6.7	7.0	1.8	1.1	0.1	0	0		4.0	20.7
		$u_{01} = 0.2$	Sel	0	1.0	24.6	**47.6**	21.1	3.7	0.1	0	0	1.9		
			Pts	0.2	3.8	6.7	7.0	1.8	1.1	0.1	0	0		4.0	20.7
	C		Sel	0	1.8	34.5	**38.1**	17.1	6.2	0.7	0.2	0.1	1.3		
			Pts	0	3.9	7.9	5.9	1.7	1.2	0.2	0	0		4.0	20.8

Table 2.4 – *Continued.*

Sc.	Design			% dose declared as MTD & average patient allocation										DLT	$\bar{N}$
				$d_{i^\star 1}$ 2	$d_{i^\star 2}$ 4	$d_{i^\star 3}$ 8	$d_{i^\star 4}$ 16	$d_{i^\star 5}$ 22	$d_{i^\star 6}$ 28	$d_{i^\star 7}$ 40	$d_{i^\star 8}$ 54	$d_{i^\star 9}$ 70	None		
	D		Sel	0	0	11.6	**61.6**	26.7	0.1	0	0	0	0		
			Pts	0	3.0	4.7	10.9	2.4	0	0	0	0		4.3	21.0
	E		Sel	0	4.0	39.3	**26.8**	17.5	8.8	1.5	0.5	0.4	1.2		
			Pts	0.4	4.2	7.6	5.2	1.3	1.9	0.1	0.2	0		3.9	20.9
4			pTox	0.03	0.05	0.10	0.16	**0.25**	0.32	0.40	0.48	0.55			
	Optimal		Sel	0	0.4	6.7	23.8	**37.5**	20.7	8.5	2.1	0.3			
	A	$u_{01} = 0.6$	Sel	0	0	9.4	34.6	**34.3**	15.2	4.4	0.4	0.3	1.4		
			Pts	0.2	3.6	5.2	6.5	3.4	1.6	0.1	0.1	0		3.2	20.7
		$u_{01} = 0.2$	Sel	0	0	10.3	34.9	**36.8**	16.6	0	0	0	1.4		
			Pts	0.2	3.6	5.3	6.7	3.6	1.5	0	0	0		3.1	20.9
	B	$u_{01} = 0.6$	Sel	0	0	9.5	34.0	**35.5**	16.4	3.0	0.1	0.1	1.4		
			Pts	0.2	3.6	5.2	6.6	2.6	2.1	0.4	0	0.1		3.3	20.8
		$u_{01} = 0.2$	Sel	0	0	10.1	35.7	**37.3**	13.8	1.1	0.5	0.1	1.4		
			Pts	0.2	3.6	5.2	6.5	2.8	2.0	0.4	0	0.1		3.3	20.8
	C		Sel	1.2	0	0.3	13.3	**34.6**	27.9	20.0	1.5	0.9	0.3		
			Pts	0	3.6	6.0	5.8	2.6	2.3	0.5	0	0.1		3.4	20.9
	D		Sel	0	0	0	2.0	**43.6**	53.7	0.7	0	0	0		
			Pts	0	3.0	3.6	9.7	4.6	0.1	0	0	0		3.2	21.0
	E		Sel	0	2.0	17.1	21.6	**25.0**	20.1	9.9	1.5	1.5	1.3		
			Pts	0.4	3.8	5.5	5.2	1.7	3.1	0.3	0.6	0.1		3.4	20.7
5			pTox	0.001	0.005	0.03	0.10	0.16	**0.25**	0.38	0.50	0.60			
	Optimal		Sel	0	0	0.1	8.4	24.6	**44.2**	20.4	2.2	0.1			
	A	$u_{01} = 0.6$	Sel	0	0	0.7	12.4	38.5	**38.5**	7.8	1.8	0.3	0		
			Pts	0	3.1	3.5	5.1	4.9	3.5	0.4	0.4	0.1		2.7	21.0
		$u_{01} = 0.2$	Sel	0	0	0.8	11.8	43.2	**44.2**	0	0	0	0		
			Pts	0	3.1	3.5	5.3	5.4	3.7	0	0	0		2.5	21.0
	B	$u_{01} = 0.6$	Sel	0	0	0.7	13.6	38.8	**40.3**	6.1	0.1	0.4	0		
			Pts	0	3.1	3.5	5.3	3.6	4.3	1.1	0	0.2		2.9	21.0
		$u_{01} = 0.2$	Sel	0	0	0.7	14.8	43.5	**37.4**	2.7	0.5	0.4	0		
			Pts	0	3.1	3.5	5.2	3.7	4.2	1.0	0	0.2		2.9	20.9
	C		Sel	0	0	1.2	16.3	32.4	**41.1**	6.7	1.2	1.1	0		
			Pts	0	3.1	3.7	5.3	3.2	4.3	1.1	0.1	0.2		2.9	21.0
	D		Sel	0	0	0	12.0	81.1	**6.9**	0	0	0	0		
			Pts	0	3.0	3.0	6.8	7.8	0.4	0	0	0		2.1	21.0
	E		Sel	0	0	2.5	11.0	28.7	**35.3**	17.9	1.7	2.9	0		
			Pts	0.1	3.1	3.7	5.0	2.0	5.2	0.7	1.1	0.1		3.1	21.0
6			pTox	0.01	0.02	0.05	0.08	0.11	0.14	**0.25**	0.37	0.47			
	Optimal		Sel	0	0	0.6	3.2	7.6	15.1	**49.5**	20.6	3.4			
	A	$u_{01} = 0.6$	Sel	0	0	1.4	7.5	25.1	37.5	**20.4**	6.4	1.3	0.4		
			Pts	0	3.2	3.9	4.8	3.9	3.6	0.6	0.7	0.2		2.1	20.9
		$u_{01} = 0.2$	Sel	0	0	1.4	8.2	26.5	63.5	**0**	0	0	0.4		
			Pts	0	3.2	3.9	4.8	4.5	4.5	0	0	0		1.8	20.9
	B	$u_{01} = 0.6$	Sel	0	0	1.4	7.6	22.2	44.1	**21.9**	0.1	2.3	0.4		
			Pts	0	3.2	3.9	4.8	2.2	4.4	1.7	0.2	0.5		2.3	20.9
		$u_{01} = 0.2$	Sel	0	0	1.4	8.5	27.0	46.4	**10.6**	3.5	2.2	0.4		
			Pts	0	3.2	3.9	4.8	2.5	4.1	1.8	0.2	0.5		2.3	21.0
	C		Sel	0	0	1.1	7.4	18.0	45.7	**17.1**	5.1	5.6	0		

Table 2.4 – *Continued.*

Sc.	Design			% dose declared as MTD & average patient allocation										DLT	$\bar{N}$
				$d_{i^\star 1}$ 2	$d_{i^\star 2}$ 4	$d_{i^\star 3}$ 8	$d_{i^\star 4}$ 16	$d_{i^\star 5}$ 22	$d_{i^\star 6}$ 28	$d_{i^\star 7}$ 40	$d_{i^\star 8}$ 54	$d_{i^\star 9}$ 70	None		
			Pts	0	3.2	4.1	4.4	2.4	4.5	1.6	0.2	0.6		2.3	21.0
	D		Sel	0	0	0	9.8	73.6	16.6	**0**	0	0	0		
			Pts	0	3.0	3.1	6.4	7.8	0.7	0	0	0		1.7	21.0
	E		Sel	0	0.1	3.1	4.5	11.1	31.8	**27.9**	8.0	13.2	0.3		
			Pts	0.2	3.2	4.0	4.0	1.1	4.7	1.2	2.0	0.5		2.6	20.9
7			pTox	0.35	0.42	0.60	0.75	0.82	0.88	0.91	0.94	0.97			
	Optimal		Sel	93.9	5.7	0.4	0	0	0	0	0	0			
	A	$u_{01} = 0.6$	Sel	3.4	6.4	1.4	0	0	0	0	0	0	**88.8**		
			Pts	1.5	5.1	1.5	0.1	0	0	0	0	0		3.6	8.2
		$u_{01} = 0.2$	Sel	3.4	6.4	1.4	0	0	0	0	0	0	**88.8**		
			Pts	1.5	5.1	1.5	0.1	0	0	0	0	0		3.6	8.2
	B	$u_{01} = 0.6$	Sel	3.4	6.4	1.4	0	0	0	0	0	0	**88.8**		
			Pts	1.5	5.1	1.5	0.1	0	0	0	0	0		3.6	8.2
		$u_{01} = 0.2$	Sel	3.4	6.4	1.4	0	0	0	0	0	0	**88.8**		
			Pts	1.5	5.1	1.5	0.1	0	0	0	0	0		3.6	8.2
	C		Sel	4.1	6.9	0.7	0	0	0	0	0	0	**88.3**		
			Pts	1.0	6.1	1.5	0.1	0	0	0	0	0		3.9	8.7
	D		Sel	0.1	59.6	40.3	0	0	0	0	0	0	**0**		
			Pts	0	5.8	14.8	0.4	0	0	0	0	0		11.7	21.0
	E		Sel	6.0	5.1	0	0	0	0	0	0	0	**88.9**		
			Pts	2.1	4.6	1.0	0.1	0	0	0	0	0		3.3	7.8
8			pTox	0.001	0.005	0.01	0.02	0.04	0.05	0.10	0.16	**0.25**			
	Optimal		Sel	0.3	0	0	0	0.6	0.6	9.2	29.4	**59.9**			
	A	$u_{01} = 0.6$	Sel	0	0	0	0.3	3.3	22.6	27.9	25.6	**20.3**	0		
			Pts	0	3.1	3.1	3.3	3.5	4.1	1.1	1.5	1.3		1.1	21.0
		$u_{01} = 0.2$	Sel	0	0	0	0.3	4.3	95.4	0	0	**0**	0		
			Pts	0	3.1	3.1	3.4	3.9	7.5	0	0	0		0.6	21.0
	B	$u_{01} = 0.6$	Sel	0	0	0	0.3	3.0	24.4	38.0	0.5	**33.8**	0		
			Pts	0	3.1	3.1	3.4	0.9	4.6	2.8	0.7	2.4		1.4	21.0
		$u_{01} = 0.2$	Sel	0	0	0	0.3	4.3	30.2	20.7	11.0	**33.5**	0		
			Pts	0	3.1	3.1	3.3	1.1	4.3	3.0	0.7	2.4		1.4	21.0
	C		Sel	0	0	0	0.1	2.1	23.5	26.2	13.3	**34.8**	0		
			Pts	0	3.0	3.1	3.3	0.9	4.5	3.1	0.7	2.4		1.4	21.0
	D		Sel	0	0	0	0.3	46.2	53.5	0	0	**0**	0		
			Pts	0	3.0	3.0	3.7	9.4	1.9	0	0	0		0.6	21.0
	E		Sel	0	0	0	0.3	0.6	6.8	21.1	13.2	**58.0**	0		
			Pts	0.1	3.0	3.1	3.2	0.3	4.1	1.5	3.1	2.6		1.6	21.0

Sc.: Scenarios; **pTox**: true probability of toxicity in humans; **Sel**: proportion of times of declaring a dose as MTD; **Pts**: average number of patients allocated to a dose; $\bar{N}$: average number of patients treated per simulated trial; **None**: proportion of trials that have been stopped early without a declaration of MTD.

3

A ROBUST BAYESIAN META-ANALYTIC MODEL TO INCORPORATE PRECLINICAL ANIMAL DATA

Summary. Before a first-in-man trial is conducted, preclinical studies are performed in animals to help characterise the safety profile of the new medicine. We propose a robust Bayesian hierarchical model to synthesise animal and human toxicity data, using scaling factors to translate doses administered to different animal species onto an equivalent human scale. After scaling doses, the parameters of dose-toxicity models intrinsic to different animal species can be interpreted on a common scale. A prior distribution is specified for each translation factor to capture uncertainty about differences between toxicity of the drug in animals and humans. Information from animals can then be leveraged to learn about the relationship between dose and risk of toxicity in a new phase I trial in humans. The model allows human dose-toxicity parameters to be exchangeable with the study-specific parameters of animal species studied so far or non-exchangeable with any of them. This leads to robust inferences, enabling the model to give greatest weight to the animal data with parameters most consistent with human parameters, or discount all animal data in the case of non-exchangeability of parameters. The proposed model is illustrated using a case study and simulations. Numerical results suggest that our proposal improves the precision of estimates of the toxicity rates when animal and human data are consistent, while it discounts animal data in cases of inconsistency.

Keywords: Bayesian hierarchical model; Historical data; Oncology; Phase I clinical trials; Robustness.

3.1 INTRODUCTION

There has been much recent interest in methods leveraging historical information for the design and interpretation of new clinical trials (Viele et al. [2014]; Eichler et al. [2013, 2016]; Neuenschwander et al. [2016a]; van Rosmalen et al. [2017]). Information may be available from clinical trials, epidemiological studies, medical research or routine clinical practice. For example, patients randomised to standard of care or placebo in historical trials can be used to augment (Hobbs et al. [2013]; French et al. [2010, 2012]) or, in exceptional circumstances, substitute entirely (Eichler et al. [2016]) for the control arm of a new trial, thus enabling more ethical or smaller studies, or studies which learn more about the novel therapy. Methods for leveraging historical information have applications to trials in small or difficult to study populations, for example, paediatric trials (Wadsworth et al. [2018]) or studies of antibiotics for drug resistant pathogens (Dane and Wetherington [2014]). In the context of early phase trials, Takeda and Morita [2018] incorporate data from a completed phase I trial into a subsequent dose-escalation study performed in a different patient population. Cunanan and Koopmeiners [2017] discussed possibilities of combining information across patient populations for a more accurate characterisation of the toxicity profile of a new compound in oncology. These proposals attempt to use historical data to inform decision making when the new phase I trial data are sparse.

When leveraging historical data, it is always possible that a conflict will emerge between the historical and the new trial data. In view of this, several approaches have been developed which down-weight the historical data either to a degree that is fixed ahead of time, or determined dynamically based upon the extent of the observed prior-data conflict. Power priors by Ibrahim and Chen [2000] with a fixed exponent are examples of 'static priors' (Viele et al. [2014]) while power priors with random exponents (Duan et al. [2006]), commensurate priors (Hobbs et al. [2011, 2012]), and meta-analytic analyses based on Bayesian hierarchical random-effects models (Neuenschwander et al. [2010, 2016a]; Schmidli et al. [2014]) are examples of dynamic approaches. This manuscript will consider a meta-analytic approach to incorporate animal data from preclinical studies into a phase I oncology study.

Bayesian model-based designs for phase I dose-escalation studies in oncology use all accumulated trial data for interim decision making. These designs have shown superior operating characteristics to the traditional algorithmic 3+3 design (Storer [1989]), correctly identifying the true maximum tolerated dose (MTD) with higher probability and allocating a higher proportion of patients to this dose (Jaki et al. [2013]). So far numerous Bayesian procedures based upon one- or two-parameter models for the dose-toxicity relationship have been proposed, such as the continual reassessment method (O'Quigley et al. [1990]; Paoletti and Kramar [2009]), procedures implementing escalation with overdose control (Babb et al. [1998]), and Bayesian decision theoretic approaches which make dose recommendations to maximise a gain function (Whitehead [2006]) at interims.

Whilst a one-parameter model may provide an adequate local approximation to the dose-toxicity relationship, when linking dose-toxicity relationships in animals and humans we will find it helpful to have a more complete description of how risk varies with dose, and so adopt a two-parameter Bayesian logistic regression model (BLRM) (Whitehead and Williamson [1998]; Neuenschwander et al. [2008]). A BLRM can be implemented with either 'operational priors', so-called because they are chosen to ensure that a dose-escalation procedure has favourable operating characteristics, or priors representing substantive knowledge. Formulating informative priors for model parameters can be challenging since there may be little relevant human data to draw upon before a phase I study is performed. However, this is not to say that no relevant information will exist as phase I trials are always preceded by preclinical studies evaluating a medicine's safety profile in animals (USFDA [2005]). The question is whether, and how, we can incorporate these data into a phase I trial.

A challenge one faces when synthesising data across species is that the safe doses associated with an acceptable risk of toxicity in humans and different animal species may cover very different dosing intervals. To overcome this obstacle, we will use allometric scaling (West and Brown [2005]; Sharma and McNeill [2009]), which is a technique often used to transform an animal dose, such as the no observed adverse event level, into a human equivalent dose by adjusting for differences in size (Baker et al. [2002]). As far as we are aware, little has been written on quantitative methods for augmenting phase I trials with animal data. Instead attention has focused on using preclinical data to inform the choice of a safe starting dose for a phase I trial (USFDA [2005]; Reigner and Blesch [2001]).

The remainder of the chapter is structured as follows. In Chapter 3.2, we propose a Bayesian meta-analytic model to borrow information from one or more animal species into human trials. In Chapter 3.3, we present a case study illustrating how the proposed hierarchical model can be used to analyse animal and human data at a single analysis. In Chapter 3.4, we use examples to explore how the model can be used to leverage animal data for interim decision making in a dose-escalation trial and interprete the results of a simulation study evaluating trial operating characteristics in Chapter 3.5. Particular attention is given to evaluating the model's ability to react to a conflict between the animal data and accruing human data. We conclude in Section 6 with a discussion of possible extensions of the proposed methodology.

3.2 BAYESIAN HIERARCHICAL MODELS FOR USING HISTORICAL DATA

This section describes the ideas underlying Bayesian hierarchical models when used as a means to augment phase I clinical trials with historical data. In Chapter 3.2.1, we review standard meta-analytic models for the incorporation of historical data that are accrued from external phase I dose-escalation studies performed in humans. In Chapter 3.2.2, we propose a robust hierarchical extension to accommodate the scenario, where historical data are measurements from preclinical studies involving multiple animal species.

3.2.1 *Standard Bayesian meta-analytic models*

Let us focus on the dose-toxicity data that are routinely collected in early phase drug development, where the primary toxicity endpoint is typically dichotomous, so that a patient experiences either a dose-limiting toxicity (DLT) or no-DLT. Suppose that dose-toxicity data, denoted by $Y_1, \dots, Y_M$, are available from M historical dose-escalation studies. For $i = 1, \dots, M$, historical study i evaluated in total J_i doses, which are indexed by the discrete set of increasing doses $\mathcal{D}_i = \{d_{i1}, \dots, d_{iJ_i}\}$. Let r_{ij} and n_{ij} denote the number of subjects experiencing a DLT and the total number receiving the dose $d_{ij} \in \mathcal{D}_i$, respectively.

Throughout, we will assume there is a monotonic increasing relationship between dose and the risk of toxicity. In this setting, a two-parameter logistic regression model is commonly adopted to analyse the binary outcome data (Whitehead and Williamson [1998]; Neuenschwander et al. [2008]). Specifically, the dose-toxicity data from the ith study can be modelled as

$$\begin{aligned} r_{ij} | p_{ij}, n_{ij} \sim & \quad \text{Binomial}(p_{ij}, n_{ij}), \\ \text{logit}(p_{ij}) = & \; \theta_{1i} + \exp(\theta_{2i}) \log(d_{ij}/d_{\text{Ref}}), \\ \boldsymbol{\theta}_i | \boldsymbol{\mu}, \Psi \sim & \; \text{BVN}(\boldsymbol{\mu}, \Psi), \quad \text{for } i = 1, \dots, M, \end{aligned} \tag{3.2.1}$$

where d_{Ref} is a predefined reference dose invariant across studies and p_{ij} denotes the probability of toxicity at dose d_{ij}. With this parameterisation, $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i})$ can be handily interpreted. Namely, θ_{1i} is the log-odds of a DLT at the reference dose d_{Ref}. Model (3.2.1) comprises a "data model" as the first level and a "parameter model" as the second level. In particular, the second level of the hierarchy stipulates that $\boldsymbol{\theta}_i$s are conditionally independent samples from a common bivariate normal distribution with unknown mean μ and covariance matrix Ψ.

Parameters μ and Ψ can be inferred as part of the model-fitting process from a frequentist perspective (DerSimonian and Laird [1986]; Knapp and Hartung [2003]), while this can be quite challenging in situations when only a few historical studies are to be analysed (Gonnermann et al. [2015]; Röver et al. [2015]; Friede et al. [2017]). Alternatively, one may analyse the data by fitting a Bayesian model (Sutton and Abrams [2001]; Lunn et al. [2013]; Turner et al. [2015]). A third level will then be added to specify priors for the hyperparameters. In our parameterisation, the priors are particularly to be placed on the elements of μ and Ψ. From this point onwards, we will focus on establishing Bayesian hierarchical models for flexible borrowing of information from historical studies.

This can be thought of as a compromise between two limiting cases: (i) complete pooling of historical datasets, which occurs when the main diagonal elements of Ψ approach 0; and (ii) no borrowing of information, which occurs when the main diagonal elements of Ψ tend towards ∞. Therefore, the covariance matrix Ψ controls the degree of borrowing across all historical studies. In the Bayesian paradigm, our choice of the priors for the elements of Ψ is crucial. Gelman [2006] discuss the prior specification for variance parameters in hierarchical models. Rhodes et al. [2016] highlight choices of the prior placed on the between-study variance in Bayesian random-effects meta-analyses. We note any default priors cannot be taken for granted and must be checked.

Let us consider to incorporate historical dose-toxicity data available from humans to inform design and analysis of a new phase I clinical trial using a standard Bayesian hierarchical model. In line with the notations defined above, let $Y_{i^\star}$ and $\theta_{i^\star}$ denote the binary DLT outcomes and the parameter vector which underpins the new phase I trial, respectively. We would assume that the new parameter vector $\theta_{i^\star} = (\theta_{1i^\star}, \theta_{2i^\star})$ is exchangeable with the historical study-specific parameter vectors, $\theta_1, \ldots, \theta_M$, which lays a foundation to implement the historical borrowing. The Bayesian hierarchical model can thus accommodate data from the historical studies $i = 1, \ldots, M$ and the new study indexed by $i^\star$. The second level, say, random-effects model for the study-specific parameters now becomes $\theta_1, \ldots, \theta_M, \theta_{i^\star}|\mu, \Psi \overset{\text{i.i.d.}}{\sim} \text{BVN}(\mu, \Psi)$. Inference for $\theta_{i^\star}$ can be performed in either a *prospective* or *retrospective* manner:

(i) a meta-analytic predictive (MAP) approach quantifies *prospectively* the prior knowledge about $\theta_{i^\star}$ at the design stage of the new clinical trial $i^\star$ by the prior predictive probability density function (pdf), $f(\theta_{i^*}|Y_1,\ldots,Y_M)$, which will later be updated with the newly accrued data Y_{i^*} using Bayes Theorem.

(ii) a meta-analytic combined (MAC) analysis is *retrospective* in the sense that once the new phase I study is complete, a random-effects meta-analysis is performed to synthesise the historical data $Y_1,\ldots,Y_M$ and the new data Y_{i^*}. Beliefs about the new parameter vector θ_{i^*} are then represented by the posterior pdf, denoted by $f(\theta_{i^*}|Y_1,\ldots,Y_M,Y_{i^*})$.

It is important to note that the MAC and MAP analyses lead to equivalent results (Neuenschwander et al. [2016a]). Yet, along with the description written above, the MAP approach requires two steps for implementation. This is because $f(\theta_{i^*}|Y_1,\ldots,Y_M)$ must be represented as a prior distribution, but it cannot be derived analytically in most cases. One approach to overcome this challenge is to approximate the prior predictive pdf with a mixture of conjugate distributions, which is then taken to be the prior for θ_{i^*} so as to derive the posterior using Bayes' Theorem; see Schmidli et al. [2014] for technical details. At the design stage when only historical data are available, implementing the MAP analysis can be beneficial. However, MAC offers considerable convenience to analyse ongoing adaptive trials with accumulating data, as no approximation step is needed given the equivalence property of MAC and MAP. Without loss of generality, throughout this paper we refer to both methods as the meta-analytic (MA) approaches.

Borrowing of information from historical data is likely to offer precision gains for estimating θ_{i^*} only if the exchangeability assumption for parameters across different subgroups holds. However, this could sometimes be unrealistic in certain situations. Neuenschwander et al. [2016b] propose a robust mixture extension of MA models with partial exchangeability structures, considering the possibility that parameters may not be a priori exchangeable. Nonparametric approaches with similar motivation can be found in Leon-Novelo et al. [2012] and Müller and Mitra [2013]. We note these methods work well in most cases when the historical data are measurements on the same scale while under the risk of excessive borrowing. But, if by any chance the historical data and new data are not readily on a similar basis, even a partial exchangeability assumption of θ_is would appear to be sceptical.

One example precluding the straightforward use of standard MA models or their extensions has emerged in the field of early drug development. Specifically, when historical data are collected from preclinical studies performed in animals to inform decision making in a subsequent phase I first-in-man study, assuming the study-specific parameters to be fully or partially exchangeable may be problematic. We thus propose a flexible Bayesian MA model in the following to address additional challenges in our context.

3.2.2 *A robust model for borrowing strength across species*

Suppose that M preclinical studies have been performed in K animal species, with $K \leqslant M$, and let $\mathcal{S} = \{S_1, \ldots, S_K\}$ contain labels for the K species studied so far. Furthermore, we assume that a single animal species $\mathcal{A}_i \in \mathcal{S}$ was investigated in study i, for $i = 1, \ldots, M$. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ denote the vector listing the binary dose-limiting toxicity (DLT) outcomes (DLT or no DLT) of the n_i animals treated in study i. Finally, we suppose that the J_i doses contained in the set $\mathcal{D}_i = \{d_{i1}, \ldots, d_{iJ_i}; d_{it_1} < d_{it_2}$ for $1 \leqslant t_1 < t_2 \leqslant J_i\}$ were evaluated in study i, where r_{ij} out of n_{ij} animals that received dose d_{ij} experienced a DLT. In each study $i = 1, \ldots, M$, we assume that the risk of experiencing a DLT increases monotonically with dose and that this relationship is adequately described by a two-parameter logistic model with parameters $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i})$. Letting p_{ij} denote the DLT risk on dose d_{ij}, we model study i data as:

$$\begin{aligned} r_{ij} | p_{ij}, n_{ij} \sim & \quad \text{Binomial}(p_{ij}, n_{ij}), \quad \text{for } j = 1, \ldots, J_i \\ \text{logit}(p_{ij}) &= \theta_{1i} + \exp(\theta_{2i}) \log(\delta_{\mathcal{A}_i} d_{ij} / d_{\text{Ref}}) \\ \boldsymbol{\theta}_i | \boldsymbol{\mu}_{\mathcal{A}_i}, \Psi &\sim \text{BVN}(\boldsymbol{\mu}_{\mathcal{A}_i}, \Psi) \quad \text{with } \mathcal{A}_i \in \{S_1, \ldots, S_K\}, \end{aligned} \tag{3.2.2}$$

where d_{Ref} is a reference dose invariant across studies, defined below, and for each $k = 1, \ldots, K$,

$$\boldsymbol{\mu}_{S_k} = \begin{pmatrix} \mu_{1S_k} \\ \mu_{2S_k} \end{pmatrix} \quad \text{and} \quad \Psi = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}.$$

Variances in Ψ represent between-trial heterogeneity within an animal species. Model (3.2.2) stipulates that the intercept of the dose-toxicity model in study i is $\theta_{1i} + \exp(\theta_{2i}) \log(\delta_{\mathcal{A}_i})$, and therefore depends upon the animal species studied. For $k = 1, \ldots, K$, the term δ_{S_k} in (3.2.2) attempts to translate the doses administered to species S_k onto a common equivalent human dosing scale. After translation of animal data, similar intervals of values should characterise acceptably safe doses in each animal species and humans. Thus, θ_{1i} and θ_{2i} in (3.2.2) can be thought of, in an approximate sense, as the parameters that would have applied in study i had humans been studied rather than animal species $\mathcal{A}_i$. The translation factor δ_{S_k} reflects the relative potency of a compound in species S_k and humans; that is, if $\delta_{S_k} > 1$ ($0 < \delta_{S_k} < 1$), the same dose of a drug has a higher (lower) DLT risk in species S_k than in humans. A special case is $\delta_{S_k} = 1$, which implies a drug has a similar potency in species S_k and humans.

Now let $i^\star$ index the phase I first-in-man trial which will evaluate doses in the set $\mathcal{D}_{i^\star} = \{d_{i^\star 1}, \dots, d_{i^\star J_{i^\star}}\}$. For completeness, we refer to humans as species $\mathcal{H}$ and define the label $\mathcal{A}_{i^\star} = \mathcal{H}$, denoting that humans will be studied in the new trial. Furthermore, let $\boldsymbol{\theta}_{i^\star} = (\theta_{1i^\star}, \theta_{2i^\star})$ denote the model parameters that will underpin the new trial. We model data from study $i^\star$ as:

$$\begin{aligned} r_{i^\star j} | p_{i^\star j}, n_{i^\star j} \sim & \quad \text{Binomial}(p_{i^\star j}, n_{i^\star j}), \quad \text{for } j = 1, \dots, J_{i^\star} \\ \text{logit}(p_{i^\star j}) = & \ \theta_{1i^\star} + \exp(\theta_{2i^\star}) \log(d_{i^\star j}/d_{\text{Ref}}), \end{aligned} \tag{3.2.3}$$

where we stipulate $\delta_{\mathcal{A}_{i^\star}} = 1$ since human doses are already expressed on the common human dosing scale, and $d_{\text{Ref}} \in \mathcal{D}_{i^\star}$ is the same reference dose specified in (3.2.2).

Specification of the translation factors embedded in (3.2.2) can be informed by allometric scaling, assuming that size-related differences in drug metabolism and pharmacokinetics explain differences in DLT risk between animals and humans given the same dose. However, there will usually be uncertainty about the precise reason for differences. Treating such translation factors as random variables robustifies the borrowing of information from animals to humans in case our prior understanding of differences between species is incorrect. We propose placing a log-normal prior on each δ_{S_k}. Table 3.1 lists log-normal priors specified using information from the FDA draft guideline *Estimating the Maximum Safe Starting Dose in Initial Clinical Trials for Therapeutics in Adult Healthy Volunteers* (USFDA [2005]); details on the derivation of these priors can be found in Technical notes 3.7.1. Model(3.2.2) assumes that for each k, translation factor δ_{S_k} applies across all studies performed in species S_k since δ_{S_k} is intended to capture intrinsic differences between species S_k and humans. We may consider refining this assumption if the different studies performed in species S_k focused on distinct subgroups, e.g., mature versus juvenile animals.

Recall that if translation factors in (3.2.2) are appropriately specified, the study-specific parameters will be expressed on a common human dosing scale and there will be similarities between the population means of the $\boldsymbol{\theta}_i$s across various animal species. Assuming population means $\boldsymbol{\mu}_{S_1}, \dots, \boldsymbol{\mu}_{S_K}$ are exchangeable, we stipulate a

Table 3.1: Log-normal priors for species-specific translation factors, $\delta_{\mathcal{A}_i} \sim \mathrm{LN}(\lambda, \gamma^2)$, specified using body surface area (BSA) and body weight (BW) data documented in the FDA draft guidelines (FDA, 2005).

Species	BW (kg)		BSA (m^2)	HED in mg/kg		HED in mg/m^2	
	Reference	Working range		λ	γ	λ	γ
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
Primates:							
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

bivariate normal 'supra-species' random-effects distribution, to allow for increased borrowing of information across species, that is, for S_k, $k = 1, \ldots, K$,

$$\mu_{S_k} | \mathbf{m}, \Sigma \sim \mathrm{BVN}(\mathbf{m}, \Sigma), \qquad (3.2.4)$$

with

$$\mathbf{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \kappa\sigma_1\sigma_2 \\ \kappa\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

The random-effects distribution in (3.2.4) accounts for between-species differences in average dose-toxicity model parameters. Differences may arise due to misspecification of one or more δ_{S_k}; if there are size-dependent and size-independent differences between an animal species and humans, the latter may not be completely captured by δ_{S_k}, but can be addressed by variances in Σ.

The Bayesian hierarchical model for the preclinical data is completed by specifying prior distributions for the hyperparameters, where we implement the model setting

$$\begin{aligned} &m_1 \sim N(\nu_1, s_1^2), \quad m_2 \sim N(\nu_2, s_2^2), \\ &\tau_1 \sim HN(z_1), \quad \tau_2 \sim HN(z_2), \quad \rho \sim U(-1, 1), \\ &\sigma_1 \sim HN(c_1), \quad \sigma_2 \sim HN(c_2), \quad \kappa \sim U(-1, 1). \end{aligned} \qquad (3.2.5)$$

Here, $HN(z)$ denotes a half-normal distribution formed by truncating a $N(0, z^2)$ prior distribution to cover the interval $(0, \infty)$. Although it will not be considered here, one could allow the between-study variances in Ψ to vary across species.

We have yet to say how we relate the human study-specific parameter vector $\theta_{i^\star}$ to the animal study-specific parameters $\theta_1, \ldots, \theta_M$. We require robust borrowing of information across species, meaning that we should down-weight information from animal species with dose-toxicity model parameters dissimilar to those in humans, and discount all preclinical data if no animal species appears similar to humans. Then, for each $k = 1, \ldots, K$, we stipulate

$$\theta_{i^\star} | \mu_{S_k}, \Psi \sim \text{BVN}(\mu_{S_k}, \Psi) \quad \text{with prior probability } w_{S_k},$$

so that w_{S_k} represents the prior plausibility that $\theta_{i^\star}$ is exchangeable with the study-specific parameters in species S_k. Note that we have defined exchangeability at the level of the study-specific model parameters since $\theta_{i^\star}$ is a study-specific, rather than population mean, parameter. To robustify inferences about $\theta_{i^\star}$, we stipulate

$$\theta_{i^\star} \sim \text{BVN}(m_0, R_0) \quad \text{with prior probability } w_R,$$

where $w_R = 1 - \sum_{k=1}^{K} w_{S_k}$ is a prior non-exchangeability weight and $\text{BVN}(m_0, R_0)$ is a weakly informative prior distribution. In practice, specification of $w_{S_1}, \ldots, w_{S_K}$ will require the input of subject-matter experts such as pharmacologists or translational scientists. The robust hierarchical model is fitted using Markov chain Monte Carlo, and thus can be implemented with software such as OpenBUGS (Lunn et al. [2009]).

We note that adding a 'supra-species' level to the Bayesian hierarchical model in equation (3.2.4) allows for increased, but robust, borrowing of information across species. When all the θ_is are similar to both each other and $\theta_{i^\star}$, we can borrow strength across the related animal species to estimate the animal population mean parameters with greater precision, and thus gain additional precision for estimating $\theta_{i^\star}$. Such borrowing is robust in the sense that if we place weakly informative priors on elements of Σ and find that, say, study-specific parameters of only one animal species are similar to $\theta_{i^\star}$, posterior distributions for elements of Σ will place larger probability mass on large between-species variances. This leads to less borrowing across animal species to estimate the μ_{S_k}s, and we tend to borrow from the most relevant animal species to learn about $\theta_{i^\star}$.

Table 3.2: Ocular toxicities observed from treated patients during a phase I first-in-man trial of AUY922. Estimated risks are derived from a logistic regression model fitted to the pooled human data alone.

	Dose (mg/m^2)								
	$d_{i^\star 1}$ 2	$d_{i^\star 2}$ 4	$d_{i^\star 3}$ 8	$d_{i^\star 4}$ 16	$d_{i^\star 5}$ 22	$d_{i^\star 6}$ 28	$d_{i^\star 7}$ 40	$d_{i^\star 8}$ 54	$d_{i^\star 9}$ 70
Number of patients	3	3	4	6	11	8	16	18	24
Number of ocular AEs	0	0	0	0	0	0	0	0	2
Ocular AE risk	0.001	0.002	0.004	0.008	0.012	0.015	0.023	0.033	0.045

3.3 ILLUSTRATIVE EXAMPLE

In this section, we apply the proposed Bayesian hierarchical model to a retrospective example, synthesising preclinical and clinical ocular toxicity data on AUY922, an experimental compound intended to treat cancer (Roman et al. [2016]; Sessa et al. [2013]).

3.3.1 *Animal data*

The safety profile of AUY922 was evaluated in several preclinical studies prior to its evaluation in humans. For this compound, ocular adverse events (AEs) were thought to potentially occur in humans. Thus, the risk of this type of event was investigated in four studies performed in a total of 152 Wistar and Brown Norway rats (Roman et al. [2016]), which we will hereafter refer to as 'rats'. The ocular AE data are displayed in Figure 3.1. The first two datasets are outcomes from Studies 1 and 2 reported in Roman et al. [2016]. Since Study 1 involved male and female rats but Study 2 involved only males, we use only the male rat data from Study 1. It was not possible to extract the ocular AE data of Studies 3 and 4 from Roman et al. [2016]. Therefore, Figure 3.1 shows simulated, but plausible, data for these studies instead (slight modifications to the doses for these studies have also been made so that we will have data on various doses to fit the logistic model for rats). Data from the phase I study of AUY922 were published in Sessa et al. [2013] and are listed in Table 3.2. During the phase I trial, doses from the set $\mathcal{D}_{i^\star} = \{2, 4, 8, 16, 22, 28, 40, 54, 70\}$ mg/m^2 were available for administration. The dose-escalation study was performed according to a BLRM-guided procedure monitoring DLTs, defined as the occurrence of any clinically relevant drug-related AE or abnormal lab value. Ocular AEs were also reported separately in the clinical paper (Sessa et al. [2013]).

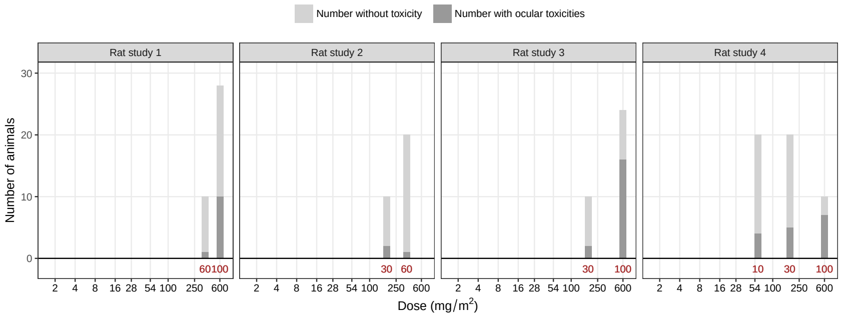

Figure 3.1: Preclinical data from four studies in rats. The height of the bar represents the number of rats studied, and the height of the dark grey segment counts the number experiencing an ocular toxicity. Doses listed in brown are the doses (mg/kg) administered to rats. Doses listed in black are the human-equivalent doses (mg/m^2). Projections are made by scaling animal doses using the prior median of δ_{Rat}.

In Chapter 3.3.2, we describe what would have been the predictive priors for the risk of an ocular AE in the phase I trial given the rat data. In this example, since animal data were available from only one species, we implement the robust Bayesian hierarchical model from Chapter 3.2.2 setting $K = 1$. We note that our model can accommodate the special case that $K = 1$ if weakly informative priors are adopted for diagonal elements of Σ. In Chapter 3.3.3, we refit the hierarchical model to incorporate both the rat and human data collected during the AUY922 phase I trial, and derive posterior distributions for the risk of an ocular AE in the human trial.

3.3.2 *Predictive priors for the risk of ocular toxicity in humans*

Setting $d_{\text{Ref}} = 28$ mg/m^2, we use the four rat datasets to fit the hierarchical model proposed in Chapter 3.2 with the following priors. We set $m_1 \sim N(-1.099, 1.98^2)$ which implies a 95% prior credible interval for the risk of toxicity at 28 mg/m^2 is 0.007 to 0.942 and prior median 0.250. Furthermore, we set $m_2 \sim N(0, 0.99^2)$ to permit flat to very steep dose-toxicity curves. These are weakly informative priors that place probability mass on plausible values of the model parameters Gelman et al. [2008]. A similar approach is used to specify the parameters of the $\text{BVN}(\mathbf{m}_0, R_0)$ non-exchangeability prior. For the variance parameters, we set $\tau_1 \sim HN(0.5)$ assuming substantial variability between the study-specific θ_{i1}s, and $\tau_2 \sim HN(0.25)$, assuming a smaller degree of variability between the slopes of study-specific dose-toxicity curves. Larger values are specified for the half-normal priors placed on σ_1 and σ_2 to preclude giving definitive information. More details are given in Technical notes 3.7.2 on the prior specification of hyperparameters. Finally, we stipulate $\delta_{\text{Rat}} \sim LN(1.792, 0.309^2)$.

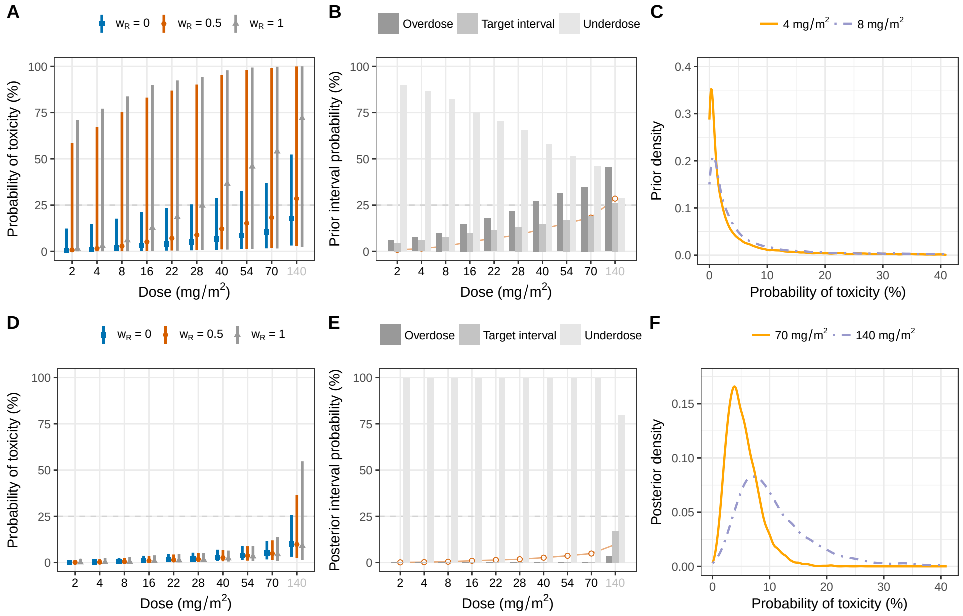

Figure 3.2: Results of the Bayesian meta-analysis, corresponding to the synthesis of ocular toxicity data in rats without and with the human data, respectively. Panels A and D show median and 95% CI of the marginal distributions for the probability of ocular toxicity. Panels B and E describe the marginal distributions of $w_R = 0.5$ using interval probabilities. The background red curve shows the median probability of toxicity of each human dose. Panels C and and F display the entire marginal distributions for the risk of ocular toxicity on doses of particular interest.

Figure 3.2A summarises predictive priors of the risk of an ocular AE in humans in the new phase I trial. Priors are derived at each human dose under a range of non-exchangeability weights. Each predictive prior is summarised by its median and 95% credible interval. Setting $w_R = 0$, predictive priors are derived assuming full exchangeability between human and animal study-specific parameters. Increasing w_R to 0.5 suggests a large degree of prior skepticism about the plausibility of such exchangeability assumption. Setting $w_R = 1$ means we discard the rat data entirely so that the prior for $\theta_{i^\star}$ is the weakly informative operational prior. Figure 3.2B further summarises priors derived setting $w_R = 0.5$ by three interval probabilities. We characterise the predictive prior for each dose by the probability: (i) of underdosing, which is said to occur if the DLT risk is less than 0.16; (ii) that the DLT risk lies in the target interval [0.16, 0.33); and (iii) of overdosing, which is said to occur if the DLT risk lies in the interval [0.33, 1] (Neuenschwander et al. [2008]). Figure 3.2C presents the predictive prior probability densities of DLT risks on two low doses, 4 and 8 mg/m^2, when $w_R = 0.5$. Such visualisations may be useful for teams to consider when selecting the starting dose for a phase I trial.

Table 3.3: Summaries of marginal predictive priors derived from the rat data setting $w_R = 0.5$. Also reported are the parameters of the Beta(a, b) approximates used for ESS calculations.

	Dose (mg/m^2)									
	d_{i^*1} 2	d_{i^*2} 4	d_{i^*3} 8	d_{i^*4} 16	d_{i^*5} 22	d_{i^*6} 28	d_{i^*7} 40	d_{i^*8} 54	d_{i^*9} 70	d_{i^*10} 140
Prior means	0.062	0.080	0.107	0.150	0.179	0.209	0.259	0.300	0.335	0.424
Prior std dev.	0.148	0.166	0.189	0.219	0.237	0.254	0.284	0.305	0.317	0.330
ESS	1.7	1.7	1.7	1.7	1.6	1.5	1.4	1.3	1.2	1.2
a	0.1	0.1	0.2	0.2	0.3	0.3	0.4	0.4	0.4	0.5
b	1.6	1.5	1.5	1.4	1.3	1.2	1.0	0.9	0.8	0.7

To calculate the effective sample size (ESS) (Morita et al. [2008]) of the predictive prior for the risk of an ocular AE on each human dose in the phase I human trial, we approximate each prior by a Beta(a, b) distribution with parameters chosen to match the first two moments of the prior. The ESS is then found as $(a + b)$. This follows because a Beta(a, b) prior can be thought of as representing opinion on the risk of an ocular AE after a out of $(a + b)$ patients allocated to a dose experience a toxicity, assuming nothing was known about the risk *a priori* (Zhou and Whitehead [2003]). After approximation, ESSs of predictive priors derived under $w_R = 0.5$ are listed in Table 3.3. The information represented by each prior is equivalent to that would be obtained from approximately 1.2 – 1.7 human patients, and so it is clear that there is heavy discounting of the preclinical data from 152 rats.

3.3.3 *Synthesising data on the termination of the first-in-man trial*

We now apply the proposed methodology to synthesise ocular AE data from both rats and the data from humans available on termination of the phase I human trial. Posterior distributions for the risk of an ocular AE on each human dose derived under models with different non-exchangeability weights are summarised in Figure 3.2D. Figures 3.2E-F summarise the posteriors derived setting $w_R = 0.5$.

With $w_R = 0.5$, the posterior probability of exchangeability between the rat and human study-specific parameters increases from the prior value of 0.5 to 0.82, suggesting that rat and human ocular AE data are more consistent than expected. Posterior median probabilities of an ocular AE in the human phase I study at doses 70 and 140 mg/m^2 are 0.048 (95% CI: [0.014, 0.118]) and 0.096 (95% CI: [0.025, 0.329]), respectively. These are slightly more cautious and narrower than the posterior medians and 95% CIs that would have been obtained had we discarded the rat data entirely from our inferences. Setting $w_R = 1$, the posterior median probabilities of an ocular

AE (95% CIs) at 70 mg/m^2 and 140 mg/m^2 are 0.045 [0.010, 0.137] and 0.087 [0.015, 0.558], respectively. The marginal posterior distributions of the risk of an ocular AE in the human trial at the two highest doses when $w_R = 0.5$ are shown in Figure 3.2F.

3.4 LEVERAGING ANIMAL DATA IN ADAPTIVE PHASE I CLINICAL TRIALS

In this section, we illustrate how our Bayesian hierarchical model can be used to leverage animal data for decision making in a hypothetical phase I dose-escalation trial.

3.4.1 *Trial design and determination of a safe starting dose*

Suppose a phase I dose-escalation study, labelled $i^\star$, is to be performed to estimate the MTD in humans, defined here as the dose associated with a risk of a DLT (of any type) of 25%. During the phase I trial, doses (in mg/m^2) from the set $\mathcal{D}_{i^\star} = \{2, 4, 8, 16, 22, 28, 40, 54, 70\}$ will be available for administration. We suppose that at the time of designing the dose-escalation study, three studies have been conducted in dogs. Simulated data from these hypothetical studies are presented in Figure S1 in the Web-based Supplementary Materials. In our notation, these data are represented by $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$. We analyse these data by fitting the Bayesian hierarchical model with priors setting $\tau_1 \sim \text{HN}(0.25)$ and $\tau_2 \sim \text{HN}(0.125)$, to assume moderate to small between-study variabilities for θ_{1i} and θ_{2i}, respectively, and $\delta_{\text{Dog}} \sim \text{LN}(2.996, 0.286^2)$. Priors for other parameters remain unchanged from Chapter 3.3.2.

Figure 3.3A summarises the prior predictive distributions for the DLT risk in the new human study $i^\star$ on each dose in $\mathcal{D}_{i^\star}$. Setting $w_R = 0.3$, the prior median for the DLT risk on dose 22 mg/m^2 is 0.252, with 95% CI [0.011, 0.800]. Figure 3.3B summarises these prior predictive distributions by presenting probabilities that the DLT risk lies in each of the three intervals (underdosing; target; and overdosing) defined in Section 3.2. We see that doses up to and including 16 mg/m^2 are associated with a prior predictive probability of overdosing of less than 25%. All hypothetical phase I dose-escalation studies start by allocating the first cohort 4 mg/m^2, with the possibility to de-escalate to 2 mg/m^2. On the basis of the dog data and our prior beliefs about their relevance with human data, 4 mg/m^2 appears very safe with $\mathbb{P}(p_{i^\star 2} < 0.1 \mid \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3) = 0.790$.

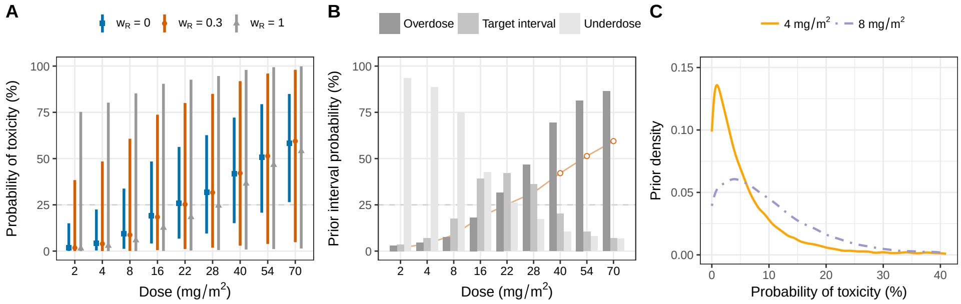

Figure 3.3: Summaries about the Bayesian analyses of the binary DLT data in dogs. Panel A shows median and 95% CI of the marginal prior predictive distribution for the probability of toxicity in the future human phase I trial, for a range of doses to be assessed. Prior predictive distributions are derived from a Bayesian meta-analysis of the dog data alone, setting w_R = 0, 0.3 or 1. Panel B gives an overview on the toxicity interval probabilities predicted based on a robust meta-analysis of dog data, setting w_R = 0.3. The background red curve shows the prior median probability of toxicity per human dose. Panel C presents prior densities for the risks of toxicity at potential starting doses.

3.4.2 *Hypothetical dose-escalation studies*

Suppose that patients enter the phase I trial in cohorts of size three and that all patients within a cohort receive the same dose. After each cohort has been treated and observed, an interim analysis is performed, at which point all dog and human data are analysed to recommend a dose for the next cohort. Cohort $h = 1$ receives 4 mg/m^2. Letting $Y_{i^\star}^{(h-1)}$ denote the vector of outcomes from the first $(h-1)$ human cohorts, the escalation rule recommends that cohort $h \geqslant 2$ receives dose

$$d_{\text{sel}}^{(h)} = \max\{d_{i^\star j} \in \mathcal{D}_{i^\star} : \mathbb{P}(p_{i^\star j} \geqslant 0.33 | \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \mathbf{Y}_{i^\star}^{(h-1)}) \leqslant 0.25\}. \tag{3.4.1}$$

Dose recommendations are also subject to the additional constraint that escalation is restricted to a maximum two-fold increase in the current dose. For the dosing set considered here, this constraint implies that if the previous cohort received a dose $d_{i^\star j} \leqslant 16$ mg/m^2, the next cohort can escalate by at most one dose level so long as the overdose control criterion is satisfied.

Figure 3.4 summarises the progress of eight hypothetical phase I trials run with simulated data, which are analysed using the proposed hierarchical model setting $w_R = 0.3$. Figure 3.4A traces dose-escalation recommendations while Figure 3.4B records how the posterior probability of exchangeability between the new human and dog study-specific parameters evolves as the study progresses. For reasons of parsimony, we monitor each simulated trial until any dose is recommended for a third time.

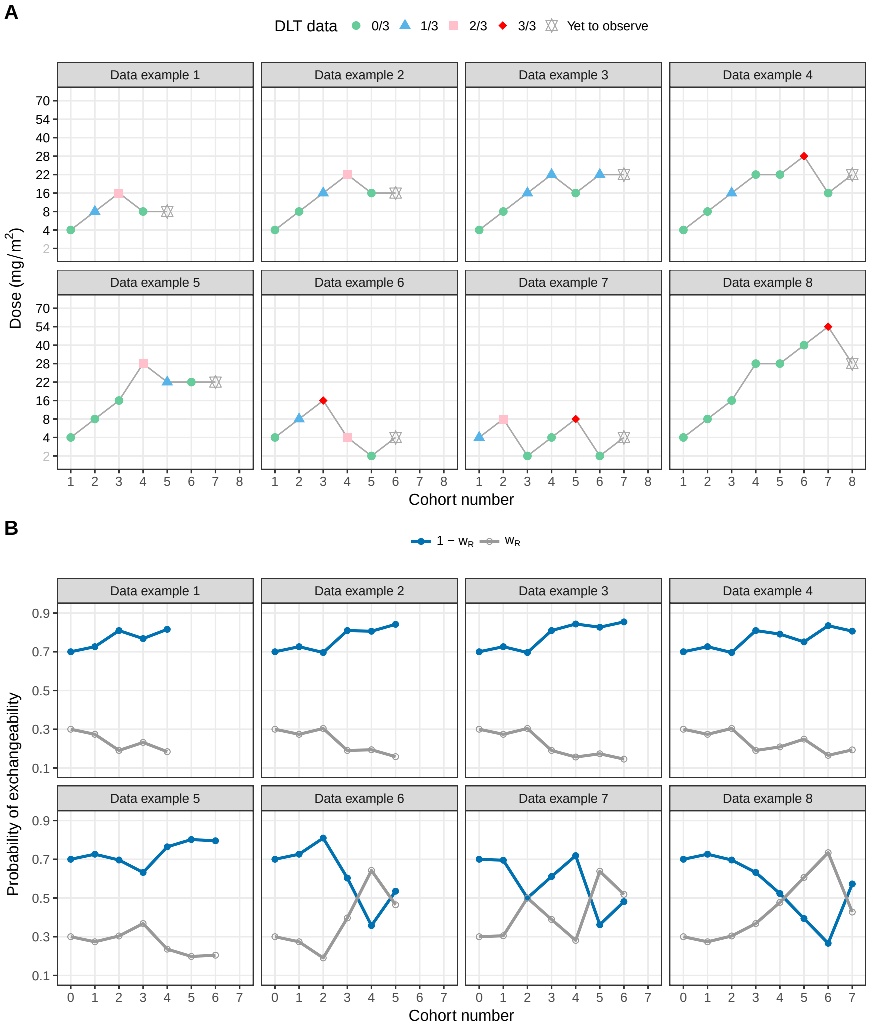

Figure 3.4: Trajectory of dose recommendations (Panel A) and posterior probabilities of exchangeability (Panel B) during the course of each hypothetical phase I trial in data examples 1 to 8.

In examples 1 to 5, data were simulated so as to be largely consistent with the prior opinion illustrated in Figure 3.3A (when $w_R = 0.3$) that the DLT risk in humans given 22 mg/m^2 in the new trial will be close to 25%, while we are confident that the risks of toxicity on 2, 4 and 8 mg/m^2 will all be well below 33%. This consistency leads to higher posterior exchangeability probabilities, as shown in Figure 3.4B. In contrast, examples 6 to 8 represent cases where there is a conflict between the human data and what was anticipated based on the analysis of the dog data.

In examples 6 and 7, the simulated human data appear consistent with a higher DLT risk at lower doses than what was predicted *a priori*. In example 6, one out of three patients in the second cohort treated with 8 mg/m^2 are observed with a DLT; we escalated to administer 16 mg/m^2 to the third cohort and all three patients experienced a DLT. Preclinical data from dog studies were then discounted, with a drop in the posterior probability of exchangeability from 0.810 to 0.358. A similar response to early observations of DLTs on low doses was seen in example 7.

In example 8, the first DLT was observed only after dosing reached 54 mg/m^2, so that the DLT risk at high doses appeared to be lower than what was predicted on the basis of the dog data. This prior-data conflict resulted in the posterior probability of exchangeability shifting from its prior value of 0.7 to 0.266 once data were available from the first six cohorts. Since the prior predictive distribution derived from the dog data suggested that the human MTD in the new study would likely lie in the neighbourhood of 22 mg/m^2, it is not surprising that dose escalation slowed down as we approached this dosing range. After completion of the forth cohort, posterior probabilities of overdose at doses 28 and 40 mg/m^2 were 0.085 and 0.293, respectively. Thus, despite the fact that no human DLTs had been observed, the procedure repeated administration of 28 mg/m^2 to the fifth cohort.

We would like to add one more note here regarding the coherence concerns in the dose-escalation procedure. Specifically, a de-escalation (an escalation) of dose is said to be coherent only when the previous patient does (does not) experience a DLT (Cheung [2005, 2011]). This can be understood as an allocation restriction for more ethical escalations and de-escalations that

$$\mathbb{P}(d_{\text{sel}}^{(h)} - d_{\text{sel}}^{(h-1)} > 0 | Q^{(h-1)} \geqslant q) = 0, \tag{3.4.2}$$

where $Q^{(h-1)}$ denote the number of DLTs observed from the $(h-1)$-th patient cohort, and q is the threshold over which the escalation would be prohibited, as well as that

$$\mathbb{P}(d_{\text{sel}}^{(h)} - d_{\text{sel}}^{(h-1)} < 0 | Q^{(h-1)} < q) = 0. \tag{3.4.3}$$

Escalation after observing one out of three patients to experience a DLT in examples 1 – 4 and 6, as was shown on Figure 3.4, may be considered as coherence violations when setting $q = 1$. In dose-escalation procedures that use our hierarchical model for leveraging preclinical animal data, this is most likely to happen when allocating a very small w_R, the prior probability of non-exchangeability, in the presence of severe inconsistency between preclinical and clinical data. One may carefully calibrate the start-off value for w_R to maintain coherence in dose escalation and de-escalation. Alternatively, a more diffuse prior distribution may be considered on the variance parameters, especially τ_1 and τ_2, which take account of heterogeneity between the standardised animal studies and the phase I first-in-man trial. This would essentially be a case-by-case issue: we recommend investigators who are interested in using our Bayesian model to evaluate their dose-escalation procedures in terms of the coherence property before applying them in a real trial.

3.5 SIMULATION STUDY

We performed a simulation study to evaluate the operating characteristics of a phase I dose-escalation procedure. We simulate trials which proceed sequentially, recruiting patients in cohorts of size three. Trials proceed using the Bayesian hierarchical model of Chapter 3.2 to leverage the dog data illustrated in Figure S1. The preclinical data are held fixed in the analysis of all simulated trials. At each analysis, we fit the Bayesian hierarchical model with four choices for w_R:

- Model A: Full exchangeability between the θ_is and $\theta_{i^\star}$ ($w_R = 0$);
- Model B: High level of prior confidence in the exchangeability assumption ($w_R = 0.3$);
- Model C: Prior ambivalence about the exchangeability assumption ($w_R = 0.5$);
- Model D: No borrowing of information from the dog data ($w_R = 1$).

Interim dose recommendations are made according to rule (4.3.1), with the same caveats as described in Chapter 3.4.2. Trials end: i) once 45 patients have been treated

Table 3.4: Scenarios for the true probability of DLT in humans. For each scenario, the figure in bold indicates the target dose closest to the true MTD.

	$d_{i^\star 1}$	$d_{i^\star 2}$	$d_{i^\star 3}$	$d_{i^\star 4}$	$d_{i^\star 5}$	$d_{i^\star 6}$	$d_{i^\star 7}$	$d_{i^\star 8}$	$d_{i^\star 9}$
	2	4	8	16	22	28	40	54	70
Scenario 1	0.08	0.16	**0.25**	0.35	0.41	0.45	0.52	0.58	0.63
Scenario 2	0.01	0.04	0.11	**0.25**	0.35	0.44	0.55	0.65	0.73
Scenario 3	0.03	0.05	0.10	0.16	**0.25**	0.32	0.40	0.48	0.55
Scenario 4	0.001	0.005	0.03	0.10	0.16	**0.25**	0.38	0.50	0.60
Scenario 5	0.01	0.02	0.05	0.08	0.11	0.14	**0.25**	0.37	0.47
Scenario 6	0.003	0.006	0.01	0.02	0.05	0.08	0.15	**0.25**	0.37
Scenario 7	**0.25**	0.42	0.60	0.75	0.82	0.88	0.91	0.94	0.97
Scenario 8	0.001	0.005	0.01	0.02	0.04	0.05	0.10	0.16	**0.25**

and observed; or ii) at any interim analysis if the lowest dose would be found as excessively toxic, that is, the trial stops at interim analysis $(h-1)$ if $\mathbb{P}(p_{i^\star 1} \geqslant 0.33 \mid \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \mathbf{Y}_{i^\star}^{(h-1)}) > 0.25$. This early stopping rule has been chosen for conservatism, meaning that the first-in-man trial has to be terminated, if potentially there is a 25% chance that a patient receiving the lowest dose, d_0 which is not the starting dose, will have unacceptably high (here, defined as $\geqslant$ 33%) possibility to experience a DLT. It corresponds to the escalation rule defined in criterion (4.3.1): we would expect for a very early stopping, when the dose has been de-escalated to the lowest level and limited evidence suggests for an escalation. These two subsets of simulated trials will later be referred to as *complete* and *stopped early* trials, respectively.

We consider eight different simulation scenarios, shown in Table 3.4, for the true dose-toxicity relationship in the new phase I trial. These toxicity scenarios are not identical with those specified in Chapter 2. In particular, we chose these scenarios with respect to the predictive priors obtained from toxicity data of the available animal studies, looking towards the behaviour of our approaches when faced with a prior-data conflict. These include scenarios which are consistent with the predictive prior derived from the dog data, as well as scenarios in which the drug is more (or less) toxic than would be expected from the dog data. For each scenario and model, results are based on 2000 simulated trials.

Define $\tilde{p}_{i^\star j}$ as the point estimate (posterior median) of the DLT risk on dose $d_{i^\star j} \in \mathcal{D}_{i^\star}$. Then at the end of a *completed* trial, we estimate the MTD as:

$$\hat{d}_M = \arg \min_{d_{i\star j} \in \mathcal{D}'_{i^\star}} |\tilde{p}_{i^\star j} - 0.25|,$$

where $\mathcal{D}'_{i^\star} \subseteq \mathcal{D}_{i^\star}$ comprises all the doses that have been administered to humans during the trial and satisfy the probabilistic overdose criterion. In each simulation scenario, we record the percentage of studies which identify each dose as the MTD.

We also record the percentage of trials which *stop early* without a MTD declaration. Furthermore, averaging across the 2000 simulated trials, we report the average number of patients allocated to each dose.

Figure 3.5 compares dose-escalation procedures implemented using Models A – D in terms of the percentage of trials which correctly select the MTD (PCS), the percentage of trials which stop early for safety; and the average number of patients allocated to the true MTD. Procedures underpinned by Models B and C perform reasonably well across all eight simulation scenarios. In cases where there is a strong prior-data conflict, for example in Scenarios 6 and 8, procedures based on Model C tend to slightly outperform those based on Model B. When there is prior-data consistency, such as in Scenario 3, the relative performances are reversed, although differences between the models remain small across all scenarios.

Comparing Models B and C with Model D, we see that by leveraging the dog data we can make gains for the PCS and average number of patients assigned to the true human MTD when the dog data are predictive of DLT risks in the new phase I trial. For example, we see an increase in PCS of at least 12.9% in Scenario 3. However, Model D clearly outperforms Models B – C in Scenario 8, in terms of the average number of patients allocated to the true MTD, although smaller differences emerge in terms of the PCS.

Comparing Models B and C with Model A, we may observe the advantages of robustification in Scenarios 6 and 8, where the assumption of full exchangeability leads to underestimation of the MTD, and allocation of a higher average number of patients to lower doses. The impact of robustification when an assumption of exhangeability is appropriate is seen in Scenario 3, when PCS decreases from 55.6% ($w_R = 0$) to 45.8% ($w_R = 0.5$). In Scenario 7, Model A appears to be much more advantageous than the rest on correctly selecting the dose 2 mg/m^2 as MTD at the end of the phase I trial. Reasons for Models B – D leading to a large proportion of trials to be terminated without a declaration of MTD are largely from the conservative *early stopping* criterion, together with our choice of the starting dose to be 4 mg/m^2 for the phase I clinical trials. Most of the simulated trials, implemented based on Models B – D, are stopped after administration of 4 mg/m^2, which is an overly toxic dose. In particular, the specified (de-)escalation rule does not recommend patients in the next cohort to be treated with dose 2 mg/m^2, at which the probability of overdose is unacceptable according to our definition. Many trials are therefore *stopped early*, as there appears to be no safe dose available for administration.

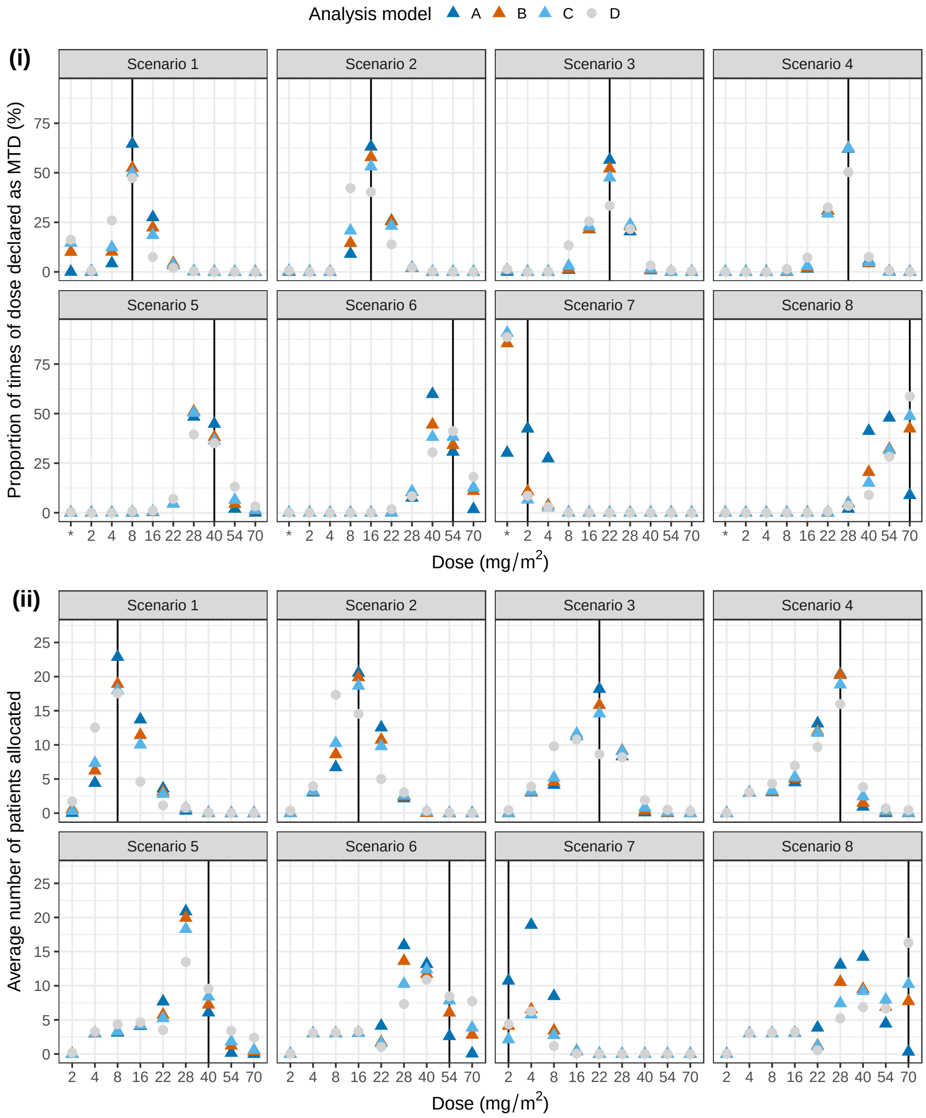

Figure 3.5: Operating characteristics of BLRM-guided dose-escalation procedures basing inferences on Models A-D, defining δ_{Dog} as a random variable. The vertical black line indicates the true MTD in humans in each simulation scenario.

For analysis Models A-C, we estimate δ_{Dog} by the median of its posterior obtained at the end of each *complete* trial. Figure 3.6 compares in each simulation scenario the distribution of posterior median estimates of δ_{Dog} with the prior median represented by the solid horizontal line. The deviation of the posterior median estimate from the prior median reflects the prior-data conflict. For example, in Scenarios 1 and 2 when preclinical data under-predict the potency of the drug in the phase I study, the posterior estimates of δ_{Dog} tend to decrease from the prior estimate to adjust for this emerging conflict. Treating δ_{Dog} as a random variable provides a mechanism to respond to prior-data conflicts and therefore further robustifies borrowing of information across species. The posterior estimates of δ_{Dog} in Scenario 7 appear to be less dispersed, because few trials were completed in this highly toxic scenario. Within a scenario, the size of the shift in posterior estimates decreases across Models A – C. As w_R increases, the need to respond to the prior-data conflict by updating δ_{Dog} becomes less as the prior weight on the exchangeability scenario decreases.

Another interesting evaluation is to compare two variants on Models A – C treating δ_{Dog} as either a random variable or a fixed constant adopted in current practice. The optimal non-parametric benchmark design (Maccario et al. [2002]) is also considered for comparison to assess potential gains of leveraging preclinical data in different simulation scenarios. Given different analysis models, we also investigated the bias, mean squared error and coverage probability of the central 95% credible interval of the posterior estimate of the DLT risk at the true MTD. Results of these assessments are available in Figures S2 and S3 in the Supplementary Materials. Furthermore, we have re-run selected simulations setting $\tau_2 \sim \text{HN}(0.25)$ instead of $\tau_2 \sim \text{HN}(0.125)$. As expected, a larger value of the scale parameter leads to reduced borrowing of information from the preclinical data while general conclusions for the comparison of different models are unchanged. Finally, we notice in practice there are situations where a phase I trial may be implemented with early stopping rules to declare the MTD. We thus consider dose-escalation procedures based on Models A – D with rules permitting early stopping when specified conditions are met. Corresponding results to demontrate trial operating characteristics are summarised in Figure S4 in the Supplementary Materials.

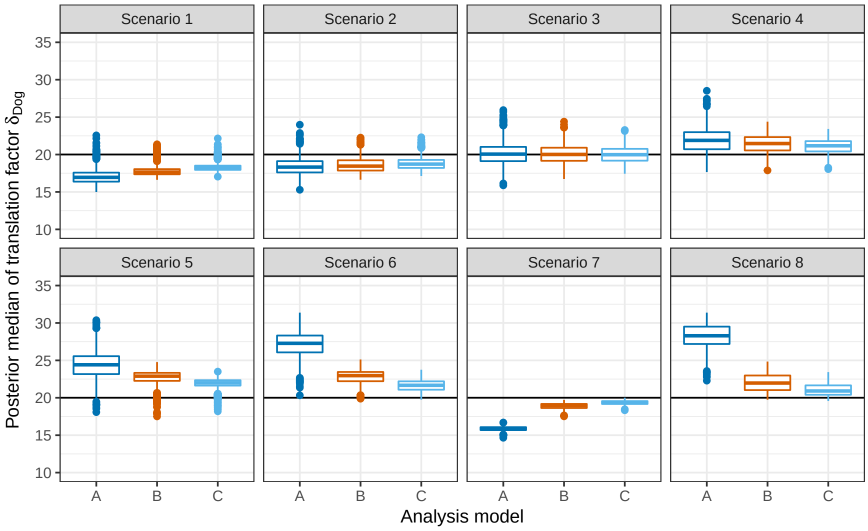

Figure 3.6: Boxplots of poesterior medians of the translation parameter δ_{Dog} under each meta-analytic model over all *complete* trials. The horizontal black line represents the prior median of δ_{Dog}.

3.6 DISCUSSION

Bayesian meta-analytic approaches provide a framework to augment a clinical trial with historical data. In this paper, we have proposed a robust Bayesian hierarchical model to augment a first-in-man trial with data from preclinical toxicology studies in animals. The simulations presented in Section 3.5 show that our methodology enables robust borrowing of information from animals to humans, and is responsive to prior-data conflicts. We note, however, that when there is a substantial prior-data conflict, using our approach may lead to a decrease in precision of the estimate, regardless of how small the prior weight assigned to the animal data is. In addition, the high proportion of trials terminated early for safety particularly in Scenario 7 was due to the conservative early stopping criterion, rather than the nature of our methodology. Investigators may relax either the bound of overdose interval or the target level, above which the trial has to be terminated if the lowest dose is thought to be overly toxic. Finding the best trade-off for correctly identifying the true MTD at the low doses, say, the least toxic two doses, without undermining the safety of patients is not the priority of our assessment. Rather, the purpose of our simulation study was to compare the proposed methodology with its alternatives that reflect opinions of (i) fully pooling the translated animal data and (ii) completely discarding relevant dose-toxicity information from the animal experiments.

Our data examples and the simulation study presented in Sections 3.3 – 3.5 have preclinical data collected from only one animal species. Additional simulations have been performed (results not reported here) to verify the performance of the meta-analytic model for cases that $K = 2$ and $K = 3$. They supported similar conclusions to those shown in this paper, namely that borrowing of information from animals to humans is robust and is led by data from the most relevant species. Having a larger number of preclinical studies involving multiple animal species is no doubt more advantageous for estimating the variance parameters that are associated with between-study and between-species heterogeneity. We also note, in a hierarchical model that assumes full exchangeability of the population means, learning about the variance parameters in the 'supra-species' level is needed to facilitate sharing of information between different species to an appropriate extent.

High quality preclinical data are essential to design an ethical phase I clinical trial (Dresser [2009]; Cook et al. [2015]). In current practice, preclinical data are used mostly to establish a safe starting dose for a phase I clinical trial. To the best of our knowledge, this paper represents a first proposal for incorporating dose-toxicity data learnt from animals into human trials. We have presented our methodology based on a two-parameter logistic regression model adopted to describe the dose-toxicity relationship. However, more sophisticated models such as physiologically based pharmacokinetic model (Gueorguieva et al. [2006]) may be considered. For the species-appropriate translation parameter introduced in our model, we assume that allometric scaling principles adjusting for body surface area (Kouno et al. [2003]; Gerina-Berzina et al. [2012]) adequately describe physiological differences between animals and humans. Additional work would be needed to verify appropriateness of this approach or refine it, since it may be inappropriate in some circumstances, for example, when the compound is a monoclonal antibody (Department of Health [2006]) or a biological agent (Tang et al. [2004]).

In this thesis, we specifically focus on the transition step from preclinical to clinical studies in early drug development, but the methodology proposed in Chapter 3.2 can be applied more broadly: it can be used to augment a clinical trial with historical data that have been recorded on a different measurement scale. Further research will extend the proposed Bayesian model to accommodate heterogeneity amongst humans. Potential applications include the case that phase I dose-escalation bridging studies to be carried out in different geographic regions. Alternatively, there may be differences between age groups, for example, between children and adults, or adults and geriatrics.

3.7 TECHNICAL NOTES

3.7.1 *Specifying a log-normal prior for the translation factor* $\delta_{\mathcal{A}_i}$

One common approach for extrapolating doses across species in practice is allometric scaling performed on the basis of body surface area (BSA). USFDA [2005] proposed calculating a human-equivalent dose (HED) by multiplying the animal dose by a factor reflecting the relationship between metabolic rate and mass in mammals:

$$\text{HED (mg/kg)} = \text{Animal dose (mg/kg)} \times \frac{(\text{BW}/\text{BSA})_{\text{Animal}}}{(\text{BW}/\text{BSA})_{\text{Human}}}, \tag{3.7.1}$$

where BW denotes the body weight (kg) and BSA is measured in square metres.

In the notation of this chapter, $\delta_{\mathcal{A}_i} = ((\text{BW}/\text{BSA})_{\text{Animal}}/(\text{BW}/\text{BSA})_{\text{Human}})$ is indeed the interspecies translation factor. As noted in Section 3.2, we fit models treating each δ_{S_k} as a random variable rather than a fixed constant to formally take account of uncertainty about translation factors. An independent log-normal prior is placed on each δ_{S_k} consistent with the translation factor in (3.7.1). Body weight is commonly modelled by a log-normal distribution, whilst for present purposes, we assume the body surface area has negligible variation in animals and humans. As both numerator and denominator of (3.7.1) are log-normally distributed, the translation factor can be described using a log-normal distribution.

Given the species-specific body weight and body surface area information available from the FDA draft guideline, displayed at the left of Table 3.1, we derive log-normal priors, based on an optimiser, so that medians and 95% CIs are in good agreement with the reference and working range of body weight. This is seen as an optimisation problem in the sense that we aim to minimise the distance between the summaries (reference and working range) and the key percentiles (namely, the 2.5th, 50th and 97.5th percentiles) of the log-normal prior. Specifically,

- For each animal species, BW/BSA can be summarised as $Q = \{q_L, q_M, q_U\}$, in which q_M corresponds to the reference value and $[q_L, q_U]$ as the limits of the working range
- The reference value is taken as median of the log-normal prior
- The log-normal variance is approximated such that the absolute distance between the implied 2.5th and 97.5th percentiles and q_L and q_U is minimised, respectively

- Likewise, derive the log-normal prior for BW/BSA in humans
- Depending on the unit of human dose, either mg/kg or mg/m^2, the log-normal prior for $\delta_{\mathcal{A}_i}$ is therefore obtained.

3.7.2 *Priors for other parameters*

Weakly informative priors for the robust component and population means $\mathbf{m}$:

- Prior for θ_{1i^*}: $m_{01} \sim N(\log\left(\frac{0.25}{1-0.25}\right), 2^2)$. This suggests that prior median for the probability of toxicity at $d_{\text{Ref}} = 28$ mg/m^2 is 0.25 and the 95% credible interval is (0.007, 0.944).
- Prior for θ_{2i^*}: $m_{02} \sim N(0, 1^2)$. This prior for the slope parameter is weakly informative as it allows for flat to very steep curves. Under this specification, when doubling the dose, the odds of a DLT is multiplied by $2^{\exp(0)} = 2$ (prior median), and the 95% credible interval for this multiplier is (1.1, 137.1).
- Priors for m_1 and m_2: $m_1 \sim N(\log\left(\frac{0.25}{1-0.25}\right), 1.98^2)$, and $m_2 \sim N(0, 0.99^2)$. These priors are similar to the ones for the robust component and therefore are also weakly informative.

Half-normal distributions are chosen for elements of the covariance matrix Ψ and Σ as follows.

- Priors for τ_1 and τ_2 that control borrowing within same species: $\tau_1 \sim HN(0.5)$, of which the key summaries, say, median and 95% credible interval, are 0.337 and (0.016, 1.121), respectively. This allows for substantial between-study heterogeneity for the intercept parameter, θ_{1i}. $\tau_2 \sim HN(0.25)$, of which the key summaries, say, median and 95% credible interval, are 0.169 and (0.008, 0.560), respectively. This allows for moderate between-study heterogeneity for the slope parameter, θ_{2i}.
- Priors for σ_1 and σ_2 that control borrowing across different animal species: $\sigma_1 \sim HN(15)$, of which the median and 95% credible interval are 10.117 and (0.470, 33.621), respectively; $\sigma_2 \sim HN(5)$, of which the median and 95% credible interval are 3.372 and (0.157, 11.207), respectively. These are diffused priors used in the paper for the special case $K = 1$.
- Priors for the correlation coefficients: $\rho \sim U(-1, 1)$ and $\kappa \sim U(-1, 1)$.

3.7.3 *OpenBUGS code to implement the robust Bayesian meta-analytic approach*

```
model{

# likelihood/sampling model
# Mdoses: total number of doses tested in animal studies
for(j in 1:Mdoses){
linA[j] <- theta[Study[j], 1]
+ exp(theta[Study[j], 2])∗log(delta[Species[j]]∗DosesA[j]/DoseRef)
logit(pToxA[j]) <- linA[j]
NtoxA[j] ~ dbin(pToxA[j], NsubA[j])
}

zero[1] <- 0
zero[2] <- 0

# theta=(theta1, theta2) derived from each animal study are ready for the use
# on the human equivalent scale
for(i in 1:Nstudy){
for(j in 1:Ndoses){
lin[i, j] <- theta[i, 1] + exp(theta[i, 2])∗log(DosesH[j]/DoseRef)
}

# theta = (theta1, theta2)
# parameters of the dose-toxicity model for each single study
# random effects for all studies
# sp.ind[i]: index function to specify
# which species the Study i belongs to
theta[i, 1] <- mu.ex.sp[sp.ind[i], 1] + re[i, 1]
theta[i, 2] <- mu.ex.sp[sp.ind[i], 2] + re[i, 2]
re[i, 1:2] ~ dmnorm(zero[1:2], prec.ex[1:2, 1:2])

  # PInd[]: matrice of the trivial/non-trivial weights
  # trivial weights for animals, no local robustification
  # to assure theta_i are fully exchangeable within same species
  sp.ind[i] ~ dcat(PInd[i, 1:(n.sp+1)])
}

# species cluster
for(k in 1:n.sp){
delta[k] <- exp(Prior.mn.delta[k] + Prior.sd.delta[k]*log.delta01[k])
log.delta01[k] ~ dnorm(0, 1)
mu.ex.sp[k, 1] <- m.ex[1] + re.m[k, 1]
mu.ex.sp[k, 2] <- m.ex[2] + re.m[k, 2]
re.m[k, 1:2] ~ dmnorm(zero[1:2], prec.sigma[1:2, 1:2])

theta.predH[k, 1] <- mu.ex.sp[k, 1] + re.h[k, 1]
theta.predH[k, 2] <- mu.ex.sp[k, 2] + re.h[k, 2]
re.h[k, 1:2] ~ dmnorm(zero[1:2], prec.ex[1:2, 1:2])
}

# default weakly-informative prior for robustification
theta.predH[(n.sp+1), 1:2] ~ dmnorm(Prior.mw[1:2], prec.sw[1:2, 1:2])
    cov.rb[1, 1] <- pow(Prior.sw[1], 2)
      cov.rb[2, 2] <- pow(Prior.sw[2], 2)
      cov.rb[1, 2] <- Prior.sw[1]*Prior.sw[2]*Prior.corr
    cov.rb[2, 1] <- cov.rb[1, 2]
      prec.sw[1:2, 1:2] <- inverse(cov.rb[1:2, 1:2])


# MA prediction
theta.star[1] <- theta.predH[which, 1]
theta.star[2] <- theta.predH[which, 2]


# wMix[]: non-trivial weights for humans to borrow strength from animals
which ~ dcat(wMix[1:(n.sp+1)])

# to monitor the exchangeability probability
# in the course of the new human trial
for(k in 1:(n.sp+1)){
```

```
prob.ex[k] <- equals(which, k)
}


# human data
for(j in 1:Ndoses){
linH[j] <- theta.star[1] + exp(theta.star[2])*log(DosesH[j]/DoseRef)
logit(pToxH[j]) <- linH[j]
NtoxH[j] ~ dbin(pToxH[j], NsubH[j])

pCat[j, 1] <- step(pTox.cut[1] - pToxH[j])
pCat[j, 2] <- step(pTox.cut[2] - pToxH[j])
- step(pTox.cut[1] - pToxH[j])
pCat[j, 3] <- step(1 - pToxH[j])
- step(pTox.cut[2] - pToxH[j])

}

# priors: Prior.mt1, Prior.mt2
prec.mt1 <- pow(Prior.mt1[2], -2)
prec.mt2 <- pow(Prior.mt2[2], -2)

# numerical stability:
# constrained to -10 and +10 (mt1), -5 and 5 (mt2)
m.ex[1] ~ dnorm(Prior.mt1[1], prec.mt1)I(-10, 10)
m.ex[2] ~ dnorm(Prior.mt2[1], prec.mt2)I(-5, 5)


# Priors for hyper parameters of the covariance matrix prec.ex[1:2, 1:2]
prec.tau1 <- pow(Prior.tau.HN[1], -2)
prec.tau2 <- pow(Prior.tau.HN[2], -2)
tau[1] ~ dnorm(0, prec.tau1)I(0.001,)
tau[2] ~ dnorm(0, prec.tau2)I(0.001,)
cov.ex[1, 1] <- pow(tau[1], 2)
cov.ex[2, 2] <- pow(tau[2], 2)
cov.ex[1, 2] <- tau[1]*tau[2]*rho
cov.ex[2, 1] <- cov.ex[1, 2]
prec.ex[1:2, 1:2] <- inverse(cov.ex[1:2, 1:2])

rho ~ dunif(Prior.rho[1], Prior.rho[2])

# Priors for hyper parameters of the covariance matrix prec.sigma[1:2, 1:2]
prec.sigma1 <- pow(Prior.sigma.HN[1], -2)
prec.sigma2 <- pow(Prior.sigma.HN[2], -2)
sigma[1] ~ dnorm(0, prec.sigma1)I(0.001,)
sigma[2] ~ dnorm(0, prec.sigma2)I(0.001,)
cov.sig[1, 1] <- pow(sigma[1], 2)
cov.sig[2, 2] <- pow(sigma[2], 2)
cov.sig[1, 2] <- sigma[1]*sigma[2]*kappa
cov.sig[2, 1] <- cov.sig[1, 2]
prec.sigma[1:2, 1:2] <- inverse(cov.sig[1:2, 1:2])

kappa ~ dunif(Prior.kappa[1], Prior.kappa[2])

}
```

3.8 supplementary materials

3.8.1 *The hypothetical dog data*

In Chapter 3.4, we suppose that prior to the phase I first-in-man trial, historical data are available from three hypothetical preclinical toxicology studies in dogs. Figure 3.7 shows the binomial data that we have used in Section 3.4, with the prior effective sample size described in Table 2.2.

Table 3.5: Summaries of marginal predictive priors derived from the dog data setting $w_R = 0.3$. Also reported are the parameters of the Beta(a, b) approximtes used for ESS calculation

	d_{i^*1}	d_{i^*2}	d_{i^*3}	d_{i^*4}	d_{i^*5}	d_{i^*6}	d_{i^*7}	d_{i^*8}	d_{i^*9}
	2	4	8	16	22	28	40	54	70
Prior means	0.057	0.085	0.135	0.223	0.282	0.338	0.431	0.509	0.573
Prior std dev.	0.118	0.133	0.153	0.176	0.186	0.195	0.212	0.222	0.226
ESS	3.0	3.4	4.0	4.6	4.9	4.8	4.4	4.1	3.8
a	0.2	0.3	0.5	1.0	1.4	1.6	1.9	2.1	2.2
b	2.8	3.1	3.5	3.6	3.5	3.2	2.5	2.0	1.6

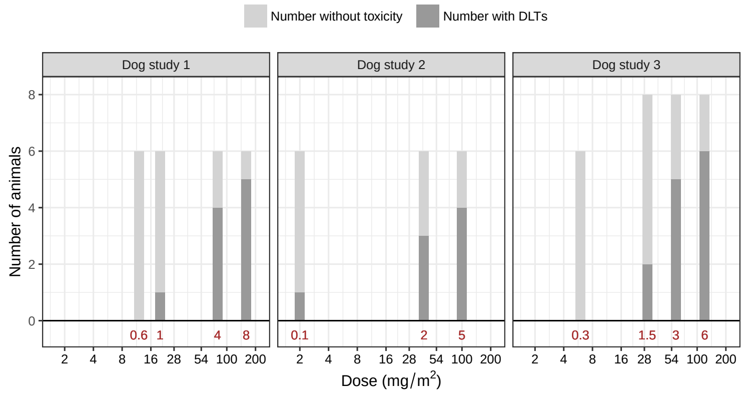

Figure 3.7: Preclinical data from three hypothetical studies in dogs. The height of the bar represents the number of dogs studied, and the height of the dark grey segment counts the number experiencing an ocular toxicity. Doses listed in brown are the doses (mg/kg) administered to dogs. Doses listed in black are the human-equivalent doses (mg/m^2). Projections are made by scaling animal doses using the prior median of δ_{Dog}.

3.8.2 *Additional simulation results*

3.8.2.1 *Numerical results of all evaluated scenarios*

The performance of trials using BLRM-guided dose-escalation under Models A – D are compared with that of the optimal non-parametric benchmark design by Maccario et al. [2002]. The optimal design is defined using the 'complete' toxicity profile of each patient, created by assuming there are $J_{i^\star}$ clones of a patient given doses spanning the dosing set $\mathcal{D}_{i^\star}$. A toxicity tolerance thereshold ϵ_n is generated from $U[0,1]$ for the nth patient, which determines the corresponding toxicity outcome at the jth dose as

$$R_{jn} = \mathbb{1}(\epsilon_n \leqslant p_{i^\star j}), \quad 1 \leqslant n \leqslant N, \quad 1 \leqslant j \leqslant J_{i^\star},$$

where $\mathbb{1}(\cdot)$ is the indicator function. An unbiased estimate for $p_{i^\star j}$ is thus $\bar{R}_j(N) = \frac{1}{N}\sum_{n=1}^{N} R_{jn}$ for a trial of which the maximum sample size is N. Consequently, the estimated MTD under the benchmark design is

$$\hat{d}_M^{\text{opt}} = \arg \min_{j=1,\ldots,J_{i^\star}} |\bar{R}_j(N) - 0.25|.$$

Improvements beyond this bound are generally not possible unless strong parametric assumptions are made about dose-response relationships. In our context, we wish to quantify the gains that can be made over the benchmark designs, in part due to borrowing strength from the preclinical data.

Two variants of analysis Model A – C are evaluated, either treating the δ_{Dog} as a random variable with a log-normal prior, or a fixed constant taking the median of the log-normal prior. Table 3.6 provides a complete listing of all the simulation results for analysis models defined in Section 5 and the optimal benchmark design.

Table 3.6: Comparison of alternative analysis models in terms of the percentage of selecting a dose as MTD at the end of the trials, percentage of early stopping for safety, average patient allocation, and average number of patients with toxicity.

Sc.	Design	δ_{S_k}		% dose declared as MTD & average patient allocation										DLT	$\bar{N}$
				$d_{i^\star 1}$ 2	$d_{i^\star 2}$ 4	$d_{i^\star 3}$ 8	$d_{i^\star 4}$ 16	$d_{i^\star 5}$ 22	$d_{i^\star 6}$ 28	$d_{i^\star 7}$ 40	$d_{i^\star 8}$ 54	$d_{i^\star 9}$ 70	None		
1			pTox	0.08	0.16	**0.25**	0.35	0.41	0.45	0.52	0.58	0.63			
	Optimal		Sel	0.4	19.2	**58.7**	19.3	1.9	0.4	0.1	0	0			
	Model A	Par	Sel	0	4.3	**64.6**	27.6	3.3	0.2	0	0	0	0		
			Pts	0	4.4	22.9	13.7	3.6	0.4	0	0	0		12.9	45.0
		Fix	Sel	0	3.0	52.4	39.4	5.1	0.1	0	0	0	0		
			Pts	0	4.0	17.0	17.9	5.7	0.4	0	0	0		13.7	45.0
	Model B	Par	Sel	0.4	10.2	**52.4**	22.4	4.4	0.2	0	0	0	10.0		
			Pts	0.6	6.2	18.9	11.4	3.1	0.6	0	0	0		11.3	40.8
		Fix	Sel	0.4	10.9	44.8	28.8	4.7	0.2	0	0	0	10.2		
			Pts	0.6	6.2	15.9	12.6	5.0	0.6	0	0	0		11.3	40.9
	Model C	Par	Sel	0.4	12.4	**50.0**	18.6	3.6	0.3	0	0	0	14.7		
			Pts	0.3	7.3	18.0	10.0	2.8	0.6	0.1	0	0		10.6	39.1
		Fix	Sel	0.5	14.0	43.0	22.7	4.8	0.2	0	0	0	14.8		
			Pts	0.4	7.5	15.6	10.8	4.1	0.8	0	0	0		10.6	39.2
	Model D		Sel	0.6	26.0	**47.3**	7.5	1.9	0.4	0	0	0	16.3		
			Pts	1.7	12.5	17.6	4.6	1.1	0.8	0.1	0	0		9.0	38.4
2			pTox	0.01	0.04	0.11	**0.25**	0.35	0.44	0.55	0.65	0.73			
	Optimal		Sel	0	0	8.3	**70.1**	20.2	1.4	0	0	0			
	Model A	Par	Sel	0	0	9.0	**60.4**	28.1	2.5	0	0	0	0		
			Pts	0	3.0	6.7	20.6	12.5	2.2	0	0	0		11.3	45.0
		Fix	Sel	0	0	5.2	60.4	32.2	2.2	0	0	0	0		
			Pts	0	3.0	4.7	18.0	17.0	2.3	0	0	0		12.1	45.0
	Model B	Par	Sel	0	0	14.5	**56.9**	26.2	2.0	0	0	0	0.4		
			Pts	0	3.1	8.6	19.9	10.8	2.4	0	0	0		10.9	44.8
		Fix	Sel	0	0	10.9	54.9	32.1	1.7	0	0	0	0.4		
			Pts	0	3.1	7.0	17.5	14.7	2.5	0	0	0		11.6	44.8
	Model C	Par	Sel	0	0	20.9	**53.1**	23.2	2.2	0	0	0	0.6		
			Pts	0	3.2	10.3	18.6	9.8	2.6	0.2	0	0		10.0	44.7
		Fix	Sel	0	0	16.9	50.9	29.5	2.1	0	0	0	0.6		
			Pts	0	3.3	8.6	17.0	12.9	3.0	0	0	0		10.8	44.8
	Model D		Sel	0	0.4	42.2	**40.3**	13.9	2.2	0.3	0	0	0.7		
			Pts	0.4	4.0	17.3	14.6	5.0	3.0	0.3	0.1	0		9.1	44.7
3			pTox	0.03	0.05	0.10	0.16	**0.25**	0.32	0.40	0.48	0.55			
	Optimal		Sel	0	0	1.1	19.5	**50.6**	23.7	4.8	0.3	0			
	Model A	Par	Sel	0	0	1.0	19.7	**55.6**	22.9	0.8	0	0	0		
			Pts	0	3.0	4.1	11.2	18.2	8.3	0.1	0	0		10.0	45.0
		Fix	Sel	0	0	0.4	17.6	60.9	21.1	0	0	0	0		
			Pts	0	3.0	3.4	8.8	21.2	8.6	0	0	0		10.0	45.0
	Model B	Par	Sel	0	0	1.5	20.3	**51.4**	24.9	1.0	0	0	0.8		
			Pts	0	3.0	4.6	11.6	15.8	9.2	0.4	0	0		9.9	44.6
		Fix	Sel	0	0	1.2	17.3	57.6	22.8	0.3	0	0	0.8		
			Pts	0	3.0	4.0	9.0	18.8	9.7	0.2	0	0		9.9	44.7
	Model C	Par	Sel	0	0.1	1.4	22.7	**45.8**	26.0	3.4	0.1	0	0.8		
			Pts	0	3.2	4.5	9.6	17.1	9.9	0.4	0	0		9.7	44.7

Table 3.6 – *Continued.*

Sc.	Design	δ_{S_k}		% dose declared as MTD & average patient allocation										DLT	$\bar{N}$
				$d_{i^\star 1}$ 2	$d_{i^\star 2}$ 4	$d_{i^\star 3}$ 8	$d_{i^\star 4}$ 16	$d_{i^\star 5}$ 22	$d_{i^\star 6}$ 28	$d_{i^\star 7}$ 40	$d_{i^\star 8}$ 54	$d_{i^\star 9}$ 70	None		
		Fix	Sel	0	0.1	2.0	19.1	54.4	22.9	0.6	0	0	0.8		
			Pts	0	3.2	4.5	9.6	17.1	9.9	0.4	0	0		9.7	44.7
	Model D		Sel	0	0.2	13.4	25.1	**32.9**	22.4	3.3	0.9	0.4	1.4		
			Pts	0.4	3.9	9.8	10.8	8.6	8.2	1.9	0.5	0.3		8.8	44.4
4			pTox	0.001	0.005	0.03	0.10	0.16	$\underline{\textbf{0.25}}$	0.38	0.50	0.60			
	Optimal		Sel	0	0	0	1.2	22.2	**62.0**	14.4	0.2	0	0		
	Model A	Par	Sel	0	0	0	1.5	27.4	**65.6**	5.5	0	0	0		
			Pts	0	3.0	3.1	4.5	13.1	20.4	0.9	0	0		8.1	45.0
		Fix	Sel	0	0	0	0.8	33.3	65.0	0.9	0	0	0		
			Pts	0	3.0	3.0	3.7	13.5	21.7	0.1	0	0		8.1	45.0
	Model B	Par	Sel	0	0	0	1.9	29.1	**64.3**	4.3	0.4	0	0		
			Pts	0	3.0	3.1	5.0	12.0	20.3	1.5	0.1	0		8.2	45.0
		Fix	Sel	0	0	0	1.0	32.4	64.8	1.6	0.2	0	0		
			Pts	0	3.0	3.1	4.0	12.5	21.7	0.6	0	0.1		8.2	45.0
	Model C	Par	Sel	0	0	0.2	2.8	28.1	**63.8**	4.9	0.2	0	0		
			Pts	0	3.0	3.3	5.3	11.7	18.8	2.5	0.3	0.1		8.3	45.0
		Fix	Sel	0	0	0.1	2.1	30.4	64.6	2.6	0.2	0	0		
			Pts	0	3.0	3.2	4.5	11.8	21.1	1.2	0.1	0.1		8.3	45.0
	Model D		Sel	0	0	1.4	7.2	32.3	**50.6**	7.6	0.8	0.1	0		
			Pts	0	3.1	4.3	7.0	9.7	16.0	3.8	0.7	0.4		8.5	45.0
5			pTox	0.01	0.02	0.05	0.08	0.11	0.14	$\underline{\textbf{0.25}}$	0.37	0.47			
	Optimal		Sel	0	0	0	0.2	2.4	11.9	**67.2**	17.8	0.6	0		
	Model A	Par	Sel	0	0	0	0.3	4.5	48.4	**44.8**	2.0	0	0		
			Pts	0	3.0	3.1	4.1	7.7	20.9	6.1	0.2	0		5.9	45.0
		Fix	Sel	0	0	0	0.2	5.3	75.2	19.1	0.2	0	0		
			Pts	0	3.0	3.0	3.5	8.1	25.8	1.6	0	0		5.4	45.0
	Model B	Par	Sel	0	0	0.2	0.4	4.6	51.0	**38.1**	4.4	1.1	0.2		
			Pts	0	3.0	3.3	4.2	5.7	19.9	7.2	1.2	0.4		6.4	44.9
		Fix	Sel	0	0	0	0.2	5.1	62.8	25.9	4.2	1.6	0.2		
			Pts	0	3.0	3.1	3.6	6.2	23.2	4.5	0.7	0.6		6.1	44.9
	Model C	Par	Sel	0	0	0.2	0.5	4.5	50.3	**36.4**	6.5	1.4	0.2		
			Pts	0	3.0	3.3	4.2	5.2	18.3	8.4	1.8	0.6		6.7	44.8
		Fix	Sel	0	0	0	0.4	4.5	58.2	29.5	4.9	2.3	0.2		
			Pts	0	3.0	3.2	3.7	5.2	21.6	6.0	1.2	1.0		6.5	44.9
	Model D		Sel	0	0	0.7	1.1	6.9	39.6	**34.8**	13.2	3.5	0.3		
			Pts	0.2	3.3	4.4	4.7	3.5	13.5	9.6	3.4	2.4		7.7	45.0
6			pTox	0.003	0.006	0.01	0.02	0.05	0.08	0.15	$\underline{\textbf{0.25}}$	0.37			
	Optimal		Sel	0	0	0	0	0	0.4	18.3	**63.8**	17.6			
	Model A	Par	Sel	0	0	0	0	0	7.5	59.9	**30.8**	1.8	0		
			Pts	0	3.0	3.0	3.1	4.1	15.9	13.2	2.6	0.1		4.2	45.0
		Fix	Sel	0	0	0	0	0	31.6	62.3	5.2	0.9	0		
			Pts	0	3.0	3.0	3.0	4.2	24.9	6.7	0.2	0		3.4	45.0
	Model B	Par	Sel	0	0	0	0	0.2	10.3	44.5	**34.1**	10.9	0		
			Pts	0	3.0	3.0	3.1	1.7	13.6	11.8	6.0	2.8		5.4	45.0
		Fix	Sel	0	0	0	0	0.2	19.2	40.2	27.4	13.0	0		
			Pts	0	3.0	3.0	3.0	1.7	18.1	9.1	3.7	3.4		5.0	45.0
	Model C	Par	Sel	0	0	0	0	0.2	10.7	38.2	**38.3**	12.7	0		

Table 3.6 – *Continued.*

Sc.	Design	δ_{S_k}		% dose declared as MTD & average patient allocation										DLT	$\bar{N}$
				$d_{i^\star 1}$	$d_{i^\star 2}$	$d_{i^\star 3}$	$d_{i^\star 4}$	$d_{i^\star 5}$	$d_{i^\star 6}$	$d_{i^\star 7}$	$d_{i^\star 8}$	$d_{i^\star 9}$	None		
				2	4	8	16	22	28	40	54	70			
			Pts	0	3.0	3.0	3.1	1.6	10.3	12.4	7.8	3.9		6.0	45.0
		Fix	Sel	0	0	0	0	0.2	14.8	38.5	32.2	14.2	0		
			Pts	0	3.0	3.0	3.0	1.2	14.6	9.9	5.4	4.9		5.7	45.0
	Model D		Sel	0	0	0	0.2	1.8	8.3	30.4	**41.2**	18.1	0		
			Pts	0.1	3.1	3.1	3.3	1.0	7.3	10.9	8.5	7.7		7.1	45.0
7			pTox	**0.25**	0.42	0.60	0.75	0.82	0.88	0.91	0.94	0.97			
	Optimal		Sel	**90.5**	9.5	0	0	0	0	0	0	0			
	Model A	Par	Sel	**42.3**	27.5	0.1	0	0	0	0	0	0	**30.2**		
			Pts	10.7	18.9	8.5	0.4	0	0	0	0	0		16.0	38.5
		Fix	Sel	40.7	24.7	0.2	0	0	0	0	0	0	34.4		
			Pts	10.3	16.9	9.3	1.3	0	0	0	0	0		16.2	37.8
	Model B	Par	Sel	**10.8**	3.8	0.1	0	0	0	0	0	0	**85.4**		
			Pts	4.1	6.5	3.4	0.4	0	0	0	0	0		6.1	14.4
		Fix	Sel	9.8	3.4	0	0	0	0	0	0	0	86.8		
			Pts	4.1	6.3	3.4	0.4	0	0	0	0	0		5.9	14.2
	Model C	Par	Sel	**6.6**	2.6	0	0	0	0	0	0	0	**90.8**		
			Pts	2.1	5.8	2.8	0.4	0	0	0	0	0		4.9	11.1
		Fix	Sel	6.8	2.1	0	0	0	0	0	0	0	91.1		
			Pts	2.2	5.7	2.7	0.4	0	0	0	0	0		4.9	11.0
	Model D		Sel	**8.6**	2.8	0	0	0	0	0	0	0	**88.6**		
			Pts	4.3	6.3	1.2	0.1	0	0	0	0	0		4.5	11.9
8			pTox	0.001	0.005	0.01	0.02	0.04	0.05	0.10	0.16	**0.25**			
	Optimal		Sel	0	0	0	0	0	0	1.6	20.7	**77.8**			
	Model A	Par	Sel	0	0	0	0	0	2.0	41.2	47.9	**8.8**	0		
			Pts	0	3.0	3.0	3.1	3.8	13.1	14.2	4.5	0.3		3.1	45.0
		Fix	Sel	0	0	0	0	0	11.9	71.7	12.5	3.9	0		
			Pts	0	3.0	3.0	3.0	3.9	21.5	10.1	0.5	0		2.4	45.0
	Model B	Par	Sel	0	0	0	0	0.1	4.9	20.5	32.1	**42.4**	0		
			Pts	0	3.0	3.0	3.0	1.2	14.1	8.0	4.3	8.4		4.7	45.0
		Fix	Sel	0	0	0	0	0.1	7.6	21.4	26.9	44.0	0		
			Pts	0	3.0	3.0	3.0	1.2	14.1	8.0	4.3	8.4		4.4	45.0
	Model C	Par	Sel	0	0	0	0	0.1	4.6	15.0	31.5	**48.8**	0		
			Pts	0	3.0	3.0	3.1	1.1	7.4	9.2	8.0	10.2		5.3	45.0
		Fix	Sel	0	0	0	0	0.1	5.5	17.2	25.4	51.8	0		
			Pts	0	3.0	3.0	3.0	0.8	11.0	7.6	5.4	11.2		5.1	45.0
	Model D		Sel	0	0	0	0	0.6	3.5	9.0	28.2	**58.7**	0		
			Pts	0.1	3.0	3.1	3.2	0.5	5.2	6.9	6.7	16.3		6.2	45.0

Sc.: Scenario; **pTox**: true probability of toxicity in humans; **Sel**: proportion of times of declaring a dose as MTD; **Pts**: average number of patients allocated to a dose; **Par**: one variant of the meta-analytic model treating δ_{Dog} as a random variable; **Fix**: another variant of the meta-analytic model treating δ_{Dog} as a fixed constant, i.e., prior median of δ_{Dog} in our implementation.

3.8.2.2 *Improved estimation precision*

In our way of declaring a dose to be MTD, precision of the posterior estimate of the DLT risk is decisive. We thus evaluate different analysis models by examining the bias, mean squared error (MSE) and coverage probability (CP) of the central 95% credible interval of the posterior median of the DLT risk at the true MTD. Figure 3.8 visualised the comparison in terms of these metrics. As illustrated, inference based on the analysis Models A-C reports smaller bias and MSE than Model D, except the Scenarios 1 and 8.

Figure 3.9 visualised the comparison in terms of the CP of central 95% credible interval at the true MTD by applying analysis Models A-C to incorporate preclinical data versus Model D to entirely discard them. We observe that at least 95% CP was attained for almost all simulation scenarios except Scenario 1. The low convergence probability of analysis models A-C in Scenario 1 is explainable, as risk of toxicity at the dose 16 mg/m^2 tends to be underestimated and thus easier to be concluded as the MTD in humans after synthesising the dog data, which advise a safer toxicity profile of the drug to humans. Across the scenarios considered here, Model A tends to attain lowest CP for scenarios when there is a discrepancy between the prediction of human toxicity based on Bayesian meta-analysis of the dog data and the true MTD. This is because large weight placed on preclinical data would lead to excessive shrinkage to the animal parameter, although on the equivalent human dosing scale. In addition, a fixed constant of δ_{Dog} in general would produce an estimate of toxicity rate at the target dose with less accurate confidence interval.

Another interesting comparison is investigated between two variants of Bayesian meta-analytic Models A – C, by treating the translation factor δ_{Dog} as a random variable or fixed constant. As shown in both Figures 3.8 and 3.9, it is not surprising that meta-analytic models constrained with fixed δ_{Dog} would lead to increased precision in Scenarios 3 and 4 due to prior-data consistency. In other simulation scenarios, however, those models experience problems because of ineffective down-weighting of the dog data. This is especially true for Scenarios 6 and 8, when the prior predictions mismatch the true human MTD and, meanwhile, incorporating preclinical data that overestimate the human toxicity leads to more conservative dose escalations.

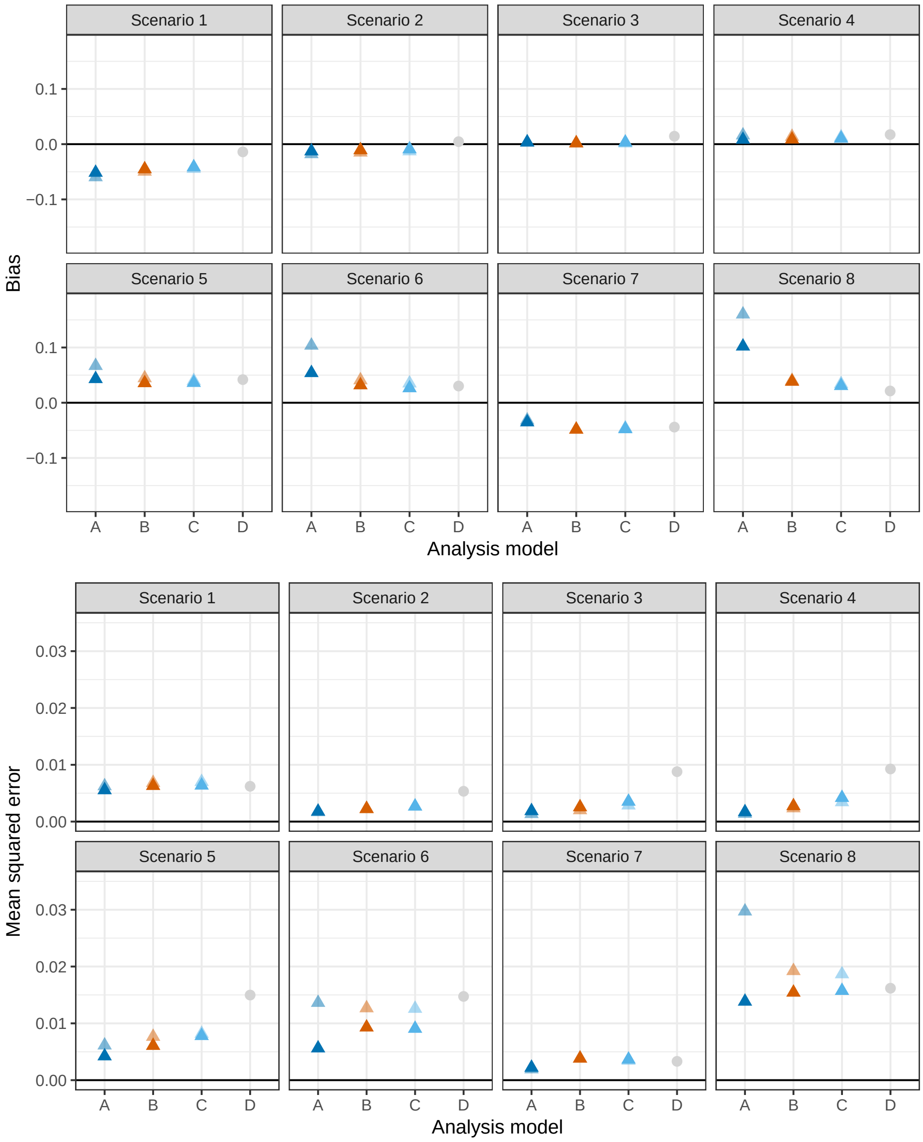

Figure 3.8: Comparison of the performance in terms of bias and mean squared error of the toxicity rate estimator at the true human MTD, based on different analysis Models A – D. Solid plotting symbols correspond to analysis models with δ_{Dog} defined as a random variable. Transparent ones correspond to the counterparts defined with δ_{Dog} as a fixed constant.

3.8.3 *Comparisons with additional early stopping rules applied*

In Figure 3.10, we show operating characteristics based on 2000 simulated phase I clinical trials under the same escalation rules, defined in Section 4.2, and the same

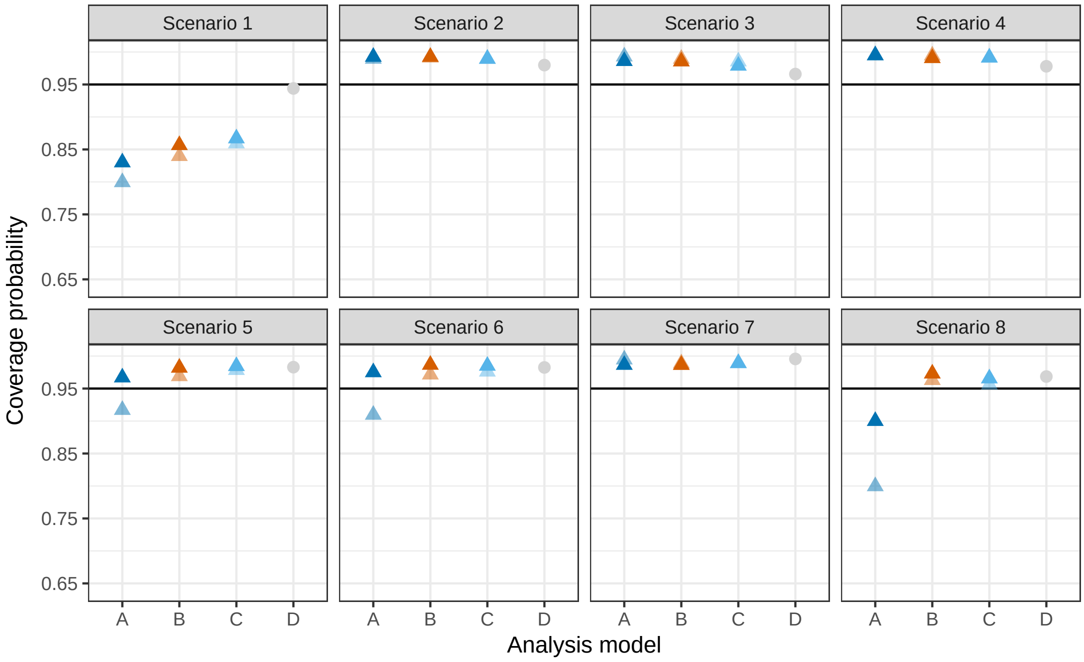

Figure 3.9: Comparison of the performance in terms of coverage probability of central 95% credible interval of the toxicity rate estimator at the true human MTD, based on different analysis Models A – D. Solid plotting symbols correspond to analysis models with δ_{Dog} defined as a random variable. Transparent ones correspond to the counterparts defined with δ_{Dog} as a fixed constant.

basis of declaring a dose to be MTD as defined in Section 5. The difference is that trials will be terminated earlier if criteria (a)-(c), specified as follows, are satisfied:

(a) There have been at least 8 cohorts (24 patients) recruited

(b) The dose selected for the next cohort is the same as the current dose

(c) At least 6 patients have been treated with this dose

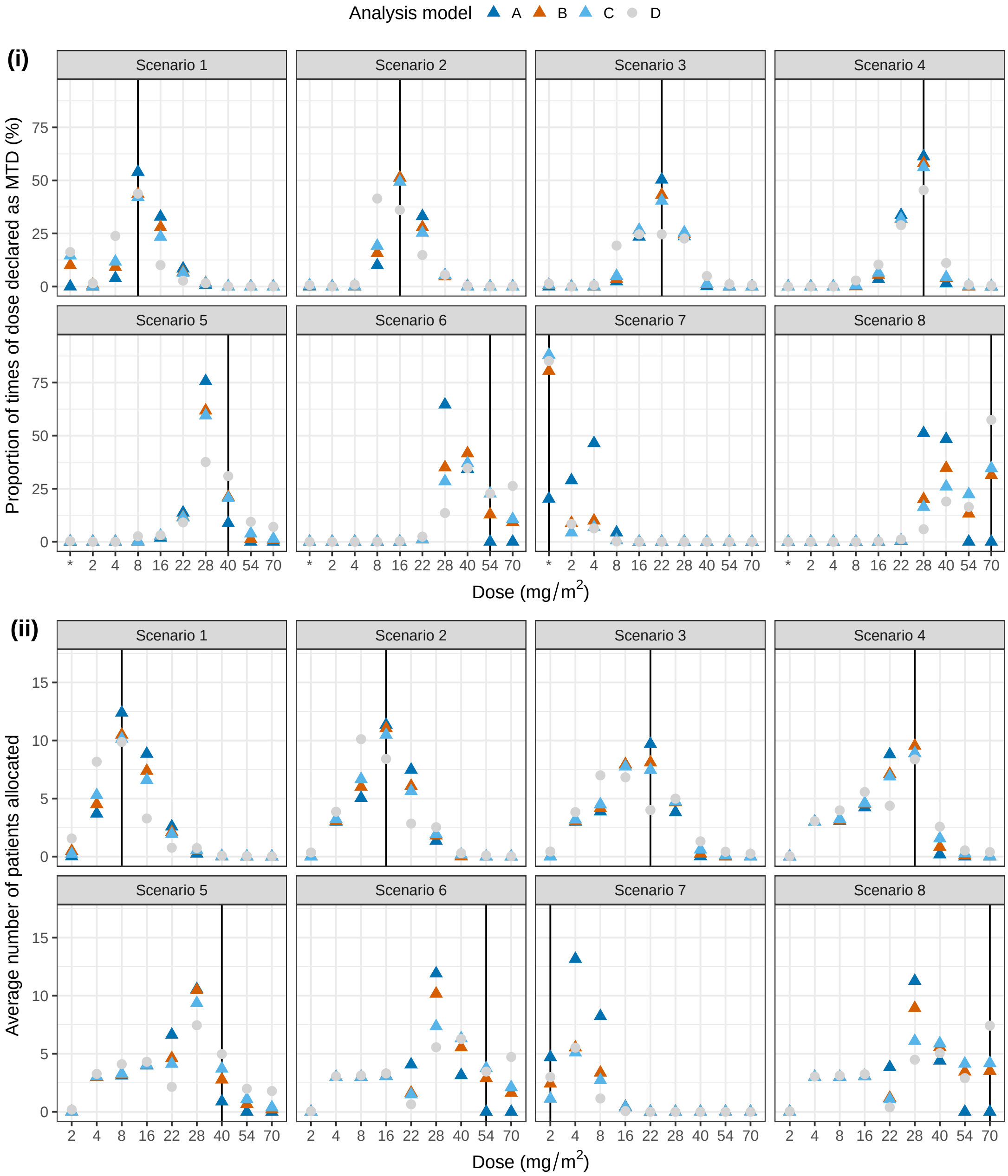

Figure 3.10: Operating characteristics of BLRM-guided dose-escalation procedures basing inferences on Models A-D, when additional early stopping rules may be applied.

4

AUGMENTING PHASE I ONCOLOGY TRIALS WITH CO-DATA: AN APPLICATION TO BRIDGING STUDIES

Summary. Phase I oncology trials may be undertaken in various geographic regions, as ethnic differences could impact on toxicity of a new medicine in distinct patient groups. This has stimulated discussions on appropriate use of external information. However, very few have looked towards incorporating preclinical animal data in such a trial setting. We fill the gap by presenting a Bayesian hierarchical model to combine information from heterogeneous sources, where intrinsic differences in toxicity of a drug between species and patient groups are addressed. Preclinical animal data are used to inform location(s) of the exchangeability distributions, from which the human dose-toxicity parameters are plausibly drawn. Our methodology is robust as it permits each dose-toxicity parameter vector that underpins a human trial to be exchangeable with similar parameter vectors or non-exchangeable with any of them, avoiding excessive shrinkage for an extreme patient group. We illustrate our Bayesian model using several representative dose-escalation trial examples in the context of bridging studies, and a simulation study to evaluate the operating characteristics. Numerical results show that our approach is responsive to both conflicts between animals and humans, and variability between patient groups.

Keywords: Bayesian hierarchical models; Bridging; Ethnic differences; Phase I clinical trials; Co-data.

4.1 INTRODUCTION

Bridging strategies are increasingly being used in the paradigm of global clinical drug development (Huang et al. [2012]; Tsong [2012]; Li and Wang [2012]; Viergever and Li [2015]), aiming to minimise duplication of clinical research without disregarding heterogeneity between patient groups. Typically, a bridging study with potentially reduced trial sample size is conducted in a new geographic region such as Japan to evaluate similarity of the performance of a medicine, which has likely been approved in other parts of the world, say, Europe, based upon a thorough drug development process from preclinical to clinical phase I-IV studies. The International Conference on Harmonisation [1998, 2006] (ICH) E5 guideline has discussed whether and when foreign data, generated from the original region to evaluate the drug, could meet the regulatory requirements of a new region where sponsors seek regitration. The amount of foreign data to be leveraged, ranging from none to full, could often be a matter of negotiations between sponsors and local regulatory authorities. Well established bridging strategies could consequently mitigate the drug lag problem (de Haen [1975]; Wileman and Mishra [2010]; Ueno et al. [2013]), and expedite supplying of new medicines to patients globally, by intergrating information from various sources, hereafter termed as co-data (Neuenschwander et al. [2016a]) to refer to all relevant, historical and concurrent, data from external studies under similar circumstances.

Over the past few decades, the Pharmaceuticals and Medical Devices Agency in Japan [2007] has been devoted to promote synchronisation of drug development with other countries. They particularly encourage domestic sponsors to participate in global clinical trials from exploratory phase I dose-finding studies, as the safety profile of a drug might be different in Caucasian and Asian patients (Morita [2011]; Ogura et al. [2014]). Both *intrinsic* factors, for example, racial genetic background, and *extrinsic* factors such as diagnostic criteria and environmental exposures, can impact on the tolerability of the drug. Mizugaki et al. [2015] reviewed 54 phase I oncology trials at the National Cancer Center Hospital between 1995 and 2012, comparing the toxicity profiles characterised based on the western and Japanese phase I clinical trials evaluating the same single agent, and have drawn a conclusion on considerable similarity in toxicity of single agent in Japan and the West.

Statistical literature has been written to discuss the impact from ethnic differences in phase I trial designs. Liu et al. [2015] develop a bridging continual reassessment method (CRM) that uses the dose-toxicity data from a completed historical trial to generate multiple sets of 'skeleton' probabilities for a new trial in another region, and apply the Bayesian model averaging approach (Yin and Yuan [2009]) to reconcile this information. Takeda and Morita [2018] present a Bayesian model-based design for a new phase I trial to use information from an external trial adaptively through a 'historical-to-current' parameter, informed by the degree of between-trial heterogeneity. In their approach, available trial data are used to formulate a suitable, but weakly informative, prior distribution at the outset of the second study; so-called weakly informative because the effective sample size (Morita et al. [2008]) of the prior is considerably small compared with the number of patients that the phase I trial will recruit.

Alternatively, relevant co-data can be generated from other relevant phase I trials which are run concurrently (or initiated in a staggered manner) to the original trial of interest, or one could leverage data on a related patient subgroup enrolled in the same trial. O'Quigley et al. [1999] propose a two-sample CRM to facilitate inferences about the maximum tolerated dose (MTD) for two distinct groups of patients simultaneously. A shift model has been further discussed in the context of bridging studies by O'Quigley and Iasonos [2014], in which the recommended dose in the second group is constrained to be the same as or several levels shifted away from what is estimated in the first group. Wages et al. [2015] extended this shift model to design a phase I/II trial of stereotactic body radiation therapy, including uncertainty that surrounds the 'true shift'.

Either performed sequentially or staggerd in time, phase I bridging trials raise an interesting research question about sharing of information between external and current trials, which may be complete or ongoing, to improve statistical inferences. One major concern is to balance available information so that decision making in a single trial, as part of the global drug development, will neither be outweighed by that made in another, nor be left to its own device for analyses. When ethnic differences have a non-negligible impact on the dose-toxicity relationship, the main objective of the phase I bridging trials would correspondingly become estimating region-specific MTDs of the drug. The goal of this research project is to propose a robust Bayesian approach that makes efficient use of available evidence, while adequately addresses heterogeneity between patient groups in phase I bridging trials.

Since no phase I clinical trials would ever be planned in a vacuum, data that preliminarily characterise the toxicity profile of a drug are commonly available from preclinical animal studies before it will be evaluated in humans (USFDA [2005]). It therefore appears to be appealing to use preclinical information, especially in above the second type of bridging studies, such that dose recommendations at early stages of the trials can also be backed up with sufficient evidence. In Chapter 3, we have discussed incorporation of animal data from multiple animal species into an ongoing phase I first-in-man trial for more ethical dose recommendations, using a Bayesian hierarchical model.

In this chapter, we propose a robust extension of their methodology to augment phase I bridging trials with co-data, which in our context consist of (i) historical animal data from complete preclinical studies before the phase I trials begin, (ii) concurrent external data from ongoing trials conducted in other geographic regions or ethnic subgroups. Figure 4.1 shows graphically what available co-data look like

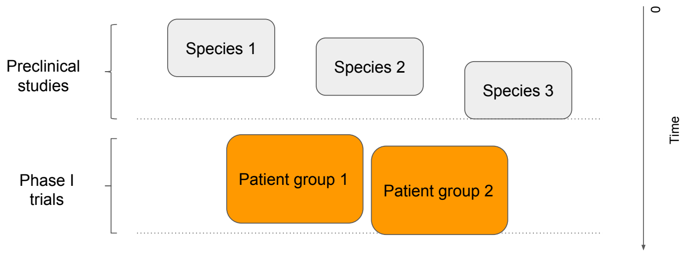

Figure 4.1: Representation of co-data for global phase I clinical trials in two patient groups.

when only two patient groups are involved. We wish to base the statistical inferences about a target dose on information from various sources, and meanwhile expect our approach to be responsive to (i) conflicts between preclinical animal data and human data from the phase I clinical trials, and (ii) possible incorrect bridging assumption that the newly generated human data from different regions may not agree with each other.

The remainder of this chapter is structured as follows. In Section 4.2, we build our work on a Bayesian hierarchical model proposed to leverage preclinical animal data into phase I trials, and discuss a robust extension to address potential heterogeneity between patient groups. In Section 4.3, we show several illustrative examples using our methodology to prospectively design global phase I trials simultaneously in two geographic regions. This is followed by a simulation study to evaluate the property of our proposal with alternative approaches used to back up statistical inferences in adaptive phase I trials in Section 4.4. In Section 4.5, we describe an application of our methodology in the context of sequential bridging studies. Finally, we discuss relevant issues raised and look towards future research in Section 4.6.

4.2 METHODS

The Bayesian hierarchical model proposed in Chapter 3 permits borrowing of information from preclinical studies in different animal species to a phase I first-in-man trial. In this section, we discuss how this random-effects model may be extended to accommodate heterogeneity between patient groups during ongoing phase I clinical trials.

Let there be a total of M preclinical studies performed in K animal species, labelled with $\mathcal{S} = \{S_1, \dots, S_K\}$. These are complete studies preceding the phase I clinical trials to be conducted in different patient groups. Indexed by $i = 1, \dots, M$, each animal study has evaluated doses d_{ij} contained in set $\mathcal{D}_i = \{d_{i1}, \dots, d_{iJ_i}; d_{it_1} \leqslant d_{it_2}$ for $1 \leqslant t_1 \leqslant t_2 \leqslant J_i\}$. On receiving a dose chosen from $\mathcal{D}_i$, outcome of each patient is recorded as either a dose-limiting toxicity (DLT) or no DLT. Let n_{ij} and r_{ij} be the number of animals treated with doses d_{ij} and the number of those experiencing a DLT, respectively. A monotone increasing in probability of toxicity, denoted by p_{ij}, along with the sorted doses d_{ij}, is assumed. We describe dose-toxicity relationship of each study i using a two-parameter logistic regression model (Whitehead and Williamson [1998]; Neuenschwander et al. [2008]), embedded with a translation parameter $\delta_{\mathcal{A}_i}$, $\mathcal{A}_i \in \mathcal{S}$:

$$\begin{aligned} r_{ij} | p_{ij}, n_{ij} &\sim \text{Binomial}(p_{ij}, n_{ij}), \text{ for } j = 1, \dots, J_i, \\ \text{logit}(p_{ij}) &= \theta_{1i} + \exp(\theta_{2i}) \log(\delta_{\mathcal{A}_i} d_{ij} / d_{\text{Ref}}), \end{aligned} \tag{4.2.1}$$

where $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i})$ are the study-specific parameters estimated on an equivalent human dosing scale, and d_{Ref} is a reference dose invariant across all dose-toxicity studies. Appropriate for each animal species, we have suggested setting log-normal priors for $\delta_{\mathcal{A}_i}$, which captures intrinsic differences between animal species S_k and humans.

Random-effects distributions are further stipulated on the second level of the meta-analytic model to facilitate borrowing of information between animal studies:

$$\boldsymbol{\theta}_i | \boldsymbol{\mu}_{\mathcal{A}_i}, \Psi \sim \text{BVN}(\boldsymbol{\mu}_{\mathcal{A}_i}, \Psi), \tag{4.2.2}$$

with

$$\boldsymbol{\mu}_{\mathcal{A}_i} = \begin{pmatrix} \mu_{1\mathcal{A}_i} \\ \mu_{2\mathcal{A}_i} \end{pmatrix} \quad \text{and} \quad \Psi = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}$$

for $\mathcal{A}_i \in \{S_1, \dots, S_K\}$. Variances in Ψ suggest between-study heterogeneity within an animal species. A 'supra-species' random effects distribution is stipulated, so as to enable increased borrowing of information between animal species. For species S_k, $k = 1, \dots, K$,

$$\boldsymbol{\mu}_{S_k} | \mathbf{m}, \Sigma \sim \text{BVN}(\mathbf{m}, \Sigma), \tag{4.2.3}$$

with

$$\mathbf{m} = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \kappa\sigma_1\sigma_2 \\ \kappa\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

The random-effects distribution in (4.2.3) specifically accouts for differences between an animal species and humans that may not be exhaustively addressed by δ_{S_k}.

Likewise, we let ℓ index the new human trials designed to evaluate doses contained in set $\mathcal{D}_\ell$. Model (4.2.1) is also applicable to describe the human toxicity data Y_ℓ, only that we may stipulate the translation parameter $\delta_{\mathcal{H}} = 1$ for the reason that doses in $\mathcal{D}_\ell$ are already expressed on the common scale; here, we have referred to humans as species $\mathcal{H}$ such that $\mathcal{A}_\ell = \mathcal{H}$. Thinking of the setting where phase I trials are to be performed in some distinct patient groups of various regions, we introduce a region parameter ϵ_ℓ into the logistic dose-toxicity model to account for potential ethnic differences. The rationale is quite the same as that for having translation factors in Model (4.2.1); that is, the dose-toxicity curve specific to a region may have its own distinct intercept. Thus, denoting the dose-toxicity parameters that underpin phase I clinical trials $\ell = 1, \ldots, L$ by $\boldsymbol{\gamma}_\ell = (\gamma_{1\ell}, \gamma_{2\ell})$,

$$\begin{aligned} r_{\ell j} | p_{\ell j}, n_{\ell j} &\sim \text{Binomial}(p_{\ell j}, n_{\ell j}), \text{ for } j = 1, \ldots, J_\ell, \\ \text{logit}(p_{\ell j}) &= \gamma_{1\ell} + \exp(\gamma_{2\ell}) \log(\epsilon_\ell d_{\ell j} / d_{\text{Ref}}), \end{aligned} \tag{4.2.4}$$

where d_{Ref} is the same reference dose used to fit Model (4.2.1).

We imagine a 'landmark' region will be chosen, which the first phase I clinical trial to launch will enroll from. Hereafter, without loss of generality, we will simplify to focus on the case $L = 2$, and label the landmark human trial as study $\mathcal{R}_L$ and the second bridging trial as study $\mathcal{R}_B$. It is then reasonable to stipulate $\epsilon_\ell = 1$ for the landmark trial $\mathcal{R}_L$, and treat any other ϵ_ℓ as a random variable in Model (4.2.4) for study $\mathcal{R}_B$. In this way, toxicity data collected on the doses used in the briding region $\mathcal{R}_B$ can be translated onto an equivalent dose-toxicity scale concerned in the landmark patient population. A prior with symmetric probability distribution may best serve our motivating problem, unless there is a strong evidence that patients from one region is more susceptible than the other. For bridging dose-finding studies, the region-specific MTDs are most likely to be the same or one dose level difference; much less frequently, the distance could be two dose levels (Liu et al. [2015]; O'Quigley and Iasonos [2014]). Distance between region-specific MTDs greater than two dose levels would cast doubt on the reliability of the assumption for bridging studies. We thus propose setting a normal prior $\epsilon_\ell \sim N(1, 0.255^2)$ truncated for non-negative values, which centres at the value of 1 and, with probability about 95%, it covers the interval (0.5, 1.5). Specifications of this region parameter ϵ_ℓ suggest, with large prior probability, that the target doses to be estimated from trials in other regions are within 0.5-fold change of that in the landmark trial(s).

Furthermore, we offer flexibility to share information between the animal studies, analysed on the common scale, and the human trials by stipulating K random-effects distributions for exchangeability (EX), as well as a weakly informative prior for non-exchangeability (NEX) to inform estimation of the human study-specific parameters $\boldsymbol{\gamma}_\ell$. This robust hierarchical extension for leveraging co-data in phase I clinical trials is in nature an EX-NEX approach, first discussed by Neuenschwander et al. [2016b] in the context of early phase clinical trials with multiple strata. Specifically, for human dose-toxicity parameters $\boldsymbol{\gamma}_\ell$, $\ell = 1, \ldots, L$, we have

(i) K EX distributions:

$$\boldsymbol{\gamma}_\ell | \boldsymbol{\mu}_{S_k}, \Phi \sim \text{BVN}(\boldsymbol{\mu}_{S_k}, \Phi), \quad \text{with prior probability } w_{\ell S_k},$$

where each of the K random-effects distributions has the unknown means consistent with those estimated by animal data of species $S_1, \ldots, S_K$, and the unknown covariance matrix $\Phi = \begin{pmatrix} \tau_3^2 & \eta\tau_3\tau_4 \\ \eta\tau_3\tau_4 & \tau_4^2 \end{pmatrix}$ specifically to describe between-trial heterogeneity of the human data;

(ii) NEX distribution:

$$\boldsymbol{\gamma}_\ell \sim \text{BVN}(\boldsymbol{m}_0, R_0), \quad \text{with prior probability } w_{\ell R},$$

where we define $w_{\ell R} = 1 - \sum w_{\ell S_k}$ for robustifying the analyses, as human parameters may be dissimilar with any of those animal parameters estimated on the common scale. Instead, they would follow their own prior $\text{BVN}(\boldsymbol{m}_0, R_0)$ independently formulated for each trial ℓ.

We will stipulate same set of prior probabilities $w_{\ell S_k}$ and $w_{\ell R}$ for all the L phase I trials at the outset, with a lack of preliminary knowledge that the parameters $\boldsymbol{\gamma}_\ell$ are more likely to follow a particular EX distribution, of which the location is informed by animal studies of species S_k. Allocating a large prior mixture weight $w_{\ell S_k}$ to an EX distribution $\text{BVN}(\boldsymbol{\mu}_{S_k}, \Phi)$ reflects a high level of prior confidence in both the bridging assumption and relevance of animal data of species S_k. We note these prior probabilities will need opinions from translational scientists or pharmacologists in practice.

There are advantages to establishing an EX-NEX model following our formulation to supplement phase I trials with co-data. We have linked the new human data with the translated animal data per species by specifying same population means $\boldsymbol{\mu}_{S_k}$, but split discussions about the between-study heterogeneity for animals and humans with Ψ and Φ, respectively. As differences between toxicity of a drug to humans and animals may not be exhaustively explained by δ_{S_k}, variances in Ψ take account of differences after such translation, suggesting how far the $\boldsymbol{\theta}_i$s are deviated from the population means that are suitable to describe the average 'true' dose-toxicity relationship on the human dosing scale. Investigators may feel more comfortable to assume the human study-specific parameters $\boldsymbol{\gamma}_\ell$s are samples from a less diffuse random-effects distribution when the bridging assumption holds. Each dose-toxicity parameter vector $\boldsymbol{\gamma}_\ell$ also has its own NEX distribution. This permits discussions on the possibility that estimates of $\boldsymbol{\gamma}_\ell$ underpinning a trial in any region might diverge from the average effects of the drug even after adjusting *intrinsic* and *extrinsic* ethnic variabilities across regions through ϵ_ℓ and Φ, respectively.

To complete our Bayesian model, we specify priors for the hyperparameters. Weakly informative priors are considered for hyperparameters of the 'supra-species' random-effects distribution that we set $m_1 \sim N(b_1, s_1^2)$ and $m_2 \sim N(b_2, s_2^2)$. In particular, parameters b_1, s_1 and b_2, s_2 will be chosen to place probability mass on plausible values of the model parameters (Gelman et al. [2008]). Priors chosen for the variance parameters should reflect opinions on the degree of heterogeneity at different levels of our model. Here, we propose setting

$$\begin{aligned} &\tau_1 \sim HN(z_1), \quad \tau_2 \sim HN(z_2), \quad \tau_3 \sim HN(z_3), \quad \tau_4 \sim HN(z_4), \\ &\sigma_1 \sim HN(c_1), \quad \sigma_2 \sim HN(c_2), \quad \rho \sim U(-1,1), \quad \kappa \sim U(-1,1), \quad \eta \sim U(-1,1), \end{aligned} \tag{4.2.5}$$

where $HN(z)$ denotes a half-normal prior distribution formed by truncating a normal distribution $N(0, z^2)$ to fall within $(0, \infty)$. This proposed robust Bayesian hierarchical model can be fitted using Markov chain Monte Carlo. We provide the OpenBUGS code in the Technical notes (Section 4.7) for implementation of our methodology.

4.3 ILLUSTRATIVE EXAMPLE

In this section, we use the proposed robust Bayesian hierarchical model to retrospectively design a trial, which aims to characterise the toxicity profile of GSK3050002 (GlaxoSmithKline [2016]), an antibody for treating patients with psoriatic arthritis. The original trial enrolled a total of 49 human subjects recruited from United Kingdom. To illustrate our approach, we suppose the evaluation was performed in the paradigm of global drug development that consist of one western trial and one eastern trial, labelled with $\mathcal{R}_L$ and $\mathcal{R}_B$ as mentioned, respectively. With this, we presume the western dose-escalation study $\mathcal{R}_L$ to be the landmark trial, and therefore have $\epsilon_\ell = 1$ for the western trial. Accordingly, for present purposes, the principal aim is now modified as estimating region-specific MTDs, defined as doses associated with a risk of DLT (of any type) of 25%. We will first show how to obtain the predictive priors for human toxicity based on animal data in Section 4.3.1, and illustrate how our robust Bayesian hierarchical model may be used to guide the dose-escalation procedure for the global phase I trials in Section 4.3.2.

4.3.1 *Hypothetical preclinical data and predictive priors for human toxicity*

We will now apply the proposed Bayesian model to a hypothetical example, for which the choice of animal species, animal doses and human doses will be informed by the set-up and background to a real phase I clinical trial evaluating safety, tolerability and pharmacokinetics of GSK3050002 (GlaxoSmithKline [2018]). Preclinical studies have been performed in monkeys and rats, among which monkeys were thought to be the most relevant animal species for predicting toxicity in humans. Animal doses 1, 10, 30, 100 mg/kg were tested in two monkey studies. The first study used four monkeys per dose group, and the second one used 10 – 12 per dose group. It was not possible to identify what dose levels were tested in rats, nor could we know the exact number of animal subjects treated and the number of toxicities observed from the trial protocol. We therefore simulated possible animal datasets according to the available but limited information. Presented in Figure 4.2, these hypothetical animal data will be used to derive predictive priors for the risk of toxicity at doses from the set $\mathcal{D}_\ell = \{0.1, 0.5, 1, 5, 10, 20\}$ mg/kg available for evaluation in both phase I clinical trials.

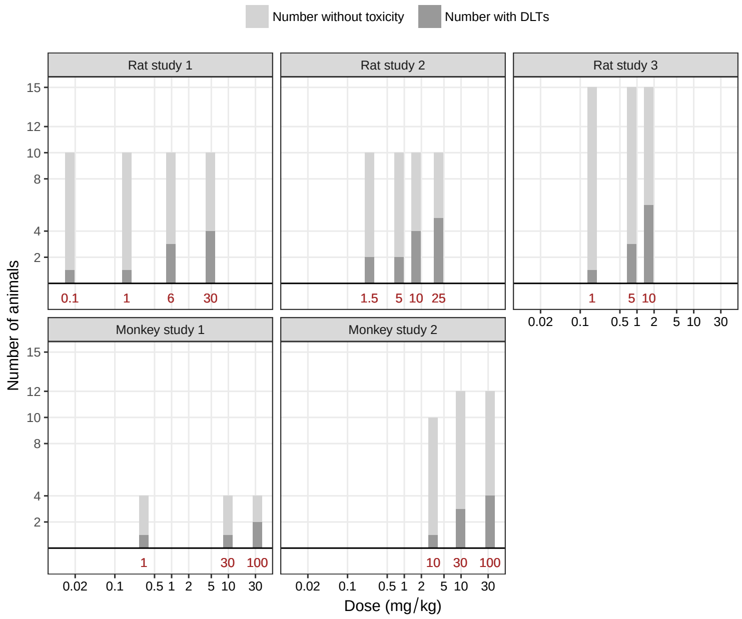

Figure 4.2: Hypothetical preclinical data in rats and monkeys. The height of the bar represents the number of animal subjects treated, and the height of the dark grey segment counts the number of toxicity. Doses listed in brown are those administered to either rats and monkeys, which are translated onto an equivalent human dosing scale in black. Projections are made by scaling animal doses using the prior median of δ_{Rat} or δ_{Monkey}.

We first derive predictive priors for human dose-toxicity parameters based on animal data of each species separately, since it will be helpful to examine whether our methodology can borrow (discount) information quickly from a particular species, which is found to be genuinely consistent (inconsistent) with the human toxicity learned from the ongoing phase I trials. Throughout the illustrative examples, we set $d_{\text{Ref}} = 10$ mg/kg and use the hypothetical preclinical datasets to fit our robust Bayesian hierarchical model proposed in Section 4.2 with the following priors. For the 'supra-species' random-effects distribution, we set $m_1 \sim N(-1.099, 1.98^2)$ and $m_2 \sim N(0, 0.99^2)$ for the global means, and stipulate $\sigma_1 \sim HN(1)$ and $\sigma_2 \sim HN(0.5)$ for variance parameters in Σ, to permit robust borrowing of information between animal species. We have considered half-normal priors $HN(z)$ with a smaller z for parameter of slope than that of intercepts, which correspond to our desire of more information to be borrowed in shapes of dose-toxicity curves compared with locations. Likewise, for the main elements in Ψ and Φ, we let $\tau_1 \sim HN(0.5), \tau_2 \sim HN(0.25)$ to take account of moderate-to-substantial heterogeneity between animal studies, and $\tau_3 \sim HN(0.25), \tau_4 \sim HN(0.125)$ of small-to-moderate heterogeneity between human groups.

Following the Bayesian meta-analytic approach elucidated in Chapter 3, we set $\delta_{\text{Rat}} \sim LN(-1.820, 0.323^2)$ and $\delta_{\text{Monkey}} \sim LN(-1.127, 0.273^2)$ to bring the animal data onto a common scale. A normal prior $\epsilon_\ell \sim N(1, 0.255^2)$ truncated to fall within (0, ∞) is placed on the region parameter. The NEX priors $BVN(\mathbf{m}_0, R_0)$ are specified for trials $\mathcal{R}_L$ and $\mathcal{R}_B$, independently, setting $m_{01} \sim N(-1.099, 2^2)$ and $m_{02} \sim N(0, 1^2)$, with a zero correlation for m_{01} and m_{02}. Predictive priors for the human dose-toxicity parameters $\boldsymbol{\gamma}_\ell$ can be obtained based on animal data from a single source of information (e.g., rat studies), by specifying the prior mixture weights, e.g., $w_{\ell\text{Rat}} = 1$, $w_{\ell\text{Monkey}} = 0$, $w_{\ell R} = 0$. When animal data are not desired to be used, we may choose $w_{\ell R} = 1$ and retain the rest as 0. Figure 4.3 show summaries of these predictive priors correspondingly.

Stand-alone analysis of animal data from a single species suggest that rat and monkey data predict doses 1 mg/kg and 5 mg/kg are likely to result in a DLT risk close to 25% when given to human subjects, respectively. Moreover, after translation of the animal doses, rat data are mainly projected on the low doses of $\mathcal{D}_\ell$. Predictive priors obtained from rat data are thus more diffuse at high doses such as 10 mg/kg and 20 mg/kg, at which monkey data in contrast have resulted in predictive priors with narrower credible intervals for the DLT risks. We may also observe that patients in trial $\mathcal{R}_B$ are predicted to have quite the same level of DLT risks compared with

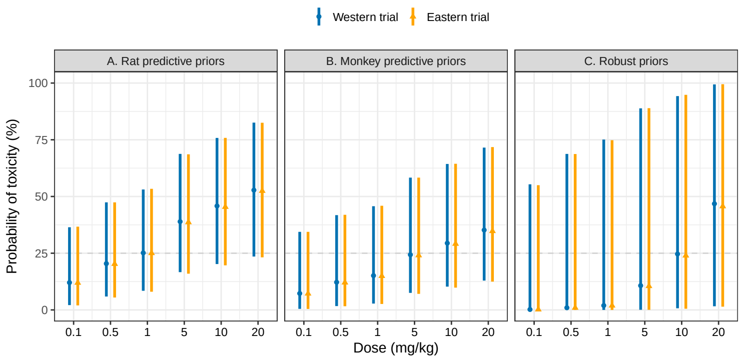

Figure 4.3: Summaries about the predictive priors for human toxicity derived based on animal data from a single species (Panels A and B) or the weakly informative prior not incorporating any animal data at all (Panel C). Medians together with 95% credible intervals of the marginal prior predictive distributions are presented.

their counterpart in trial $\mathcal{R}_L$ based on these animal data. This is because of the same prior probabilities of EX and NEX chosen for each human trial at the outset.

To leverage all available animal data, we wish to distinguish different animal species for their comparative relevance with human toxicity at the outset of trials $\mathcal{R}_L$ and $\mathcal{R}_B$. Here, we stipulate $w_{\ell \mathrm{Rat}} = 0.20$, $w_{\ell \mathrm{Monkey}} = 0.65$ and $w_{\ell \mathrm{R}} = 0.15$. Robust predictive priors are then obtained and described using medians and 95% credible intervals in Figure 4.4A. In addition, we are also interested in probabilities that a patient may be (i) underdosed, if the DLT risk is less than 0.16, (ii) correctly dosed that the DLT risk fall within the target interval [0.16, 0.33), and (iii) overdosed if the DLT risk is greater than 0.33 Neuenschwander et al. [2008]. The third interval probability of each dose is generally the most essential for recommending a safe dose to patients in the next cohort. We thus present in Figure 4.4B the prior probabilities of overdose at a range of interesting doses to be evaluated in both western and eastern trials $\mathcal{R}_L$ and $\mathcal{R}_B$. We suppose investigators would also be interested in extrapolating a safe starting dose using animal data. Presenting the prior densities of potential candidates can therefore be helpful. In this illustrative example, we show that of the lowest two doses in Figure 4.4C, and suggest choosing 0.1 mg/kg for both trials $\mathcal{R}_L$ and $\mathcal{R}_B$ to start with since this dose appears to be quite safe, based on the prior probabilities of underdose that $\mathbb{P}(p_{\ell 1} < 0.16 | \mathbf{Y}_1, \ldots, \mathbf{Y}_5) = 0.837$ for trial $\mathcal{R}_L$, and $\mathbb{P}(p_{\ell 1} < 0.16 | \mathbf{Y}_1, \ldots, \mathbf{Y}_5) = 0.840$ for trial $\mathcal{R}_B$, respectively.

As rich preclinical information may be leveraged into the phase I trials typically planned with a small sample size, we wish our EX-NEX approach is capable to discount any irrelevant animal data effectively. A fair comment on this must be based on assessment of the effective sample size (ESS) Morita et al. [2008] for each marginal

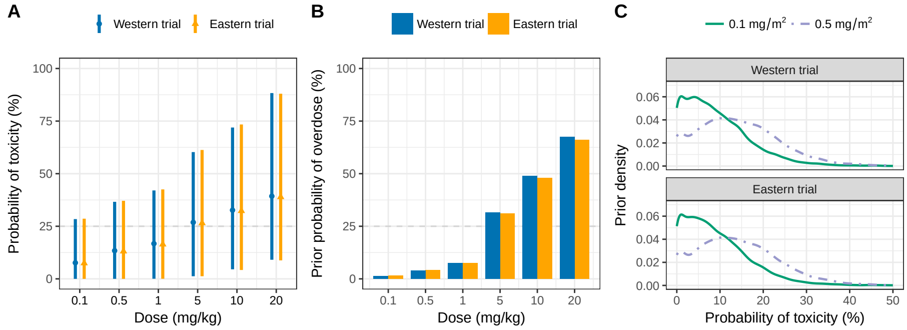

Figure 4.4: Summaries about the robust predictive priors for human toxicity based on the robust Bayesian analysis of available animal data in rats and monkeys. Panel A shows median and 95% credible interval of the marginal predictive prior for human toxicity at each dose. Panel B presents the prior interval probability of overdose, and Panel C displays prior densities for the risks of toxicity at potential starting doses.

Table 4.1: Effective sample sizes of marginal predictive priors for the DLT risk, based on robust Bayesian analysis of animal data, in the global phase I trials $\mathcal{R}_L$ and $\mathcal{R}_B$.

	Western trial, $\mathcal{R}_L$						Eastern trial, $\mathcal{R}_B$					
	d_{L1}	d_{L2}	d_{L3}	d_{L4}	d_{L5}	d_{L6}	d_{B1}	d_{B2}	d_{B3}	d_{B4}	d_{B5}	d_{B6}
	0.1	0.5	1	5	10	20	0.1	0.5	1	5	10	20
Prior means	0.093	0.145	0.177	0.280	0.343	0.416	0.093	0.144	0.175	0.278	0.341	0.413
Prior std dev.	0.084	0.104	0.114	0.143	0.158	0.179	0.086	0.106	0.116	0.145	0.160	0.180
ESS	11.0	10.5	10.2	8.9	8.1	6.6	10.4	10.0	9.7	8.6	7.8	6.5
a	1.0	1.5	1.8	2.5	2.8	2.8	1.0	1.4	1.7	2.4	2.7	2.7
b	10.0	9.0	8.4	6.4	5.3	3.9	9.4	8.6	8.0	6.2	5.1	3.8

predictive prior for DLT risk per dose, before any new data will be generated from the phase I trials. Each predictive prior for the risk of DLT in humans per dose can be approximated by beta distributions with parameters a and b, for the convenience of calculating the ESS as $(a + b)$. On deriving the Beta$(a + b)$ for both $p_{\ell j}$, we match the first two moments with the original robust predictive priors obtained using animal data. Table 4.1 lists the ESSs suggesting information represented in each marginal prior, which is thought as equivalent to what would be acquired from 6.5 – 11 human subjects treated with the doses in each trial. We note that priors are not different for the two phase I trials, as we have set same values to be taken when specifying the priors used to fit the Bayesian hierachical EX-NEX model.

4.3.2 *Design and conduct of two phase I trials in different geographic regions*

Suppose that equal trial sample size, say, 24 will be planned, and that the global drug development program will start off by recruiting patients in small cohorts of three to the western trial $\mathcal{R}_L$, which is followed by the eastern trial $\mathcal{R}_B$ with a small

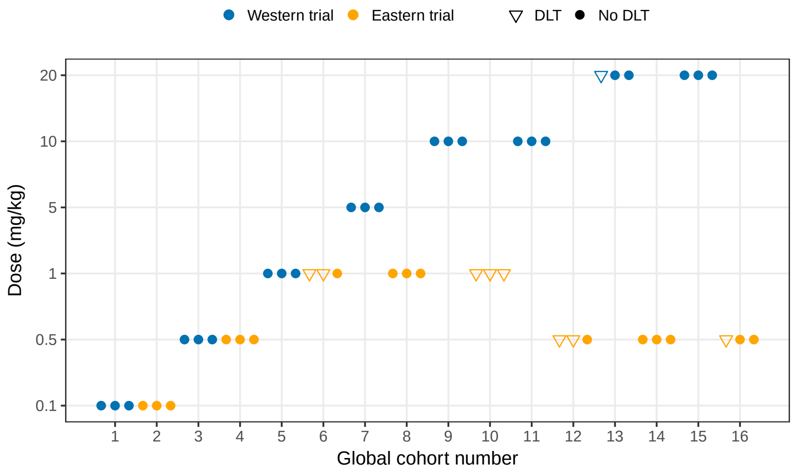

Figure 4.5: Trial trajectory of hypothetical phase I trials performed in two geographic regions, in which trial data were simulated from a divergency scenario.

delay in time. We use $h_{\ell^\star}$ and h_ℓ to index the regional cohort number of trials $\mathcal{R}_L$ and $\mathcal{R}_B$, respectively. Let the human toxicity data accumulating from the first $h_{\ell^\star}$ cohorts of trial $\mathcal{R}_L$ be $\mathbf{Y}_L^{(h_{\ell^\star})}$, and likewise their counterpart of trial $\mathcal{R}_B$ be $\mathbf{Y}_B^{(h_\ell)}$. Furthermore, we suppose that these two trials have same recruitment rate and that trial $\mathcal{R}_B$ begins after completion of the first cohort of trial $\mathcal{R}_L$. As a result, treatment of patients in both trials $\mathcal{R}_L$ and $\mathcal{R}_B$ will be undertaken in turns if integrated into one clinical drug development program, as illustrated in Figure 4.5. Each regional cohort number consequently may be recoded in the global paradigm that cohort $h_{\ell^\star}$ of the western trial becomes cohort $(2h_{\ell^\star} - 1)$ and cohort h_ℓ of the eastern trial becomes cohort $2h_\ell$. In the following, we will stick with the regional cohort number when describing inferences at interims. However, it is important to clarify that a dose to be recommended to cohort $h_{\ell^\star} \geqslant 2$ will then be based on the first $(h_{\ell^\star} - 1)$ cohorts of the western trial $\mathcal{R}_L$ and the first $(h_\ell - 1)$ cohorts, where $h_\ell \geqslant 2$, of the estern trial $\mathcal{R}_B$.

Recall that we have estimated dose 0.1 mg/kg as a suitable starting dose for patients in the first regional cohort of each trial. For $h_{\ell^\star} \geqslant 2$ and $h_\ell \geqslant 2$, a dose will be recommended for the next regional cohort according to the criterion:

$$\begin{aligned} \hat{d}_L^{(h_{\ell^\star})} &= \max\{d_{\ell j} \in \mathcal{D}_\ell : \mathbb{P}(p_{\ell j} \geqslant 0.33 | \mathbf{Y}_1, \ldots, \mathbf{Y}_5, \mathbf{Y}_L^{(h_{\ell^\star}-1)}, \mathbf{Y}_B^{(h_\ell - 1)}) \leqslant 0.25\}, \\ \hat{d}_B^{(h_\ell)} &= \max\{d_{\ell j} \in \mathcal{D}_\ell : \mathbb{P}(p_{\ell j} \geqslant 0.33 | \mathbf{Y}_1, \ldots, \mathbf{Y}_5, \mathbf{Y}_L^{(h_{\ell^\star})}, \mathbf{Y}_B^{(h_\ell - 1)}) \leqslant 0.25\}. \end{aligned} \tag{4.3.1}$$

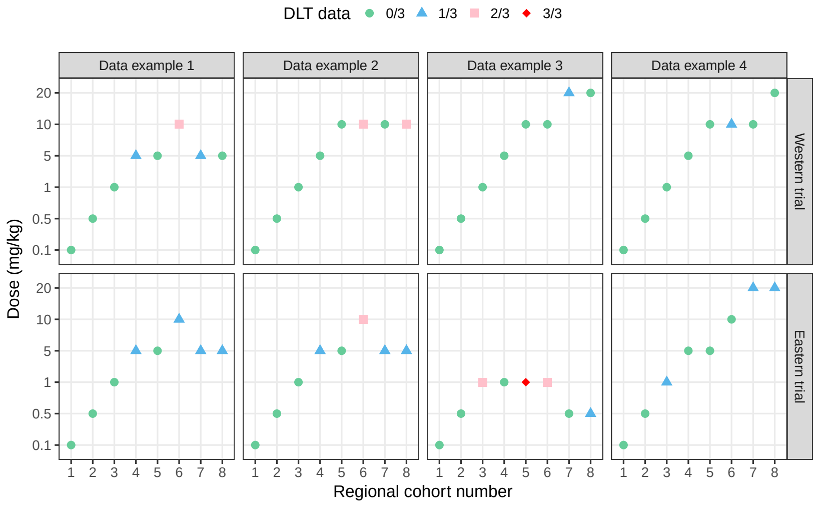

Figure 4.6: Dose-escalation scheme in each phase I clinical trial designed simultaneously using the proposed robust Bayesian hierarchical model.

To prevent too fast escalations, additional constraints such as "never skipping a dose during the escalation" may apply in practice. This means, in our illustrative example, one cannot skip dose 0.5 mg/kg to recommend 1 mg/kg for patients in cohort $h_{\ell^\star} = 2$, even the first three doses all comply with the caveat defined in (4.3.1).

We display four illustrative data examples for dose (de-)escalation of both trials $\mathcal{R}_L$ and $\mathcal{R}_B$ in Figure 4.6. Correspondingly, key summaries of the marginal posterior distributions for the DLT risk per evaluated dose on termination of each phase I trial, together with the posterior probability assigned to each of the underlying distributions of EX or NEX, are shown in Figure 4.7. These data examples were simulated under different scenarios for human toxicity and analysed using the Bayesian EX-NEX model proposed in Section 4.2. In both data examples 1 and 2, we have simulated the human toxicity data from a consistency scenario, where monkey data present very high predictability. Bridging assumption holds in all data examples, except data example 3, where considerable probability of non-exchangeability has been assigned to guide the dose-escalation procedure in the western trial $\mathcal{R}_L$, whilst the model parameters that underpin the eastern trial $\mathcal{R}_B$ have shrunk more towards the population means estimated from rat data. Data example 4 is also an interesting scenario that there exists consistency of toxicity between patient groups but conflicts between preclinical animal data and human toxicity. Using our methodology, preclinical information from rat studies, which suggests the drug to be more toxic than it actually is, was discounted substantially: as we can see by the end of each trial, the

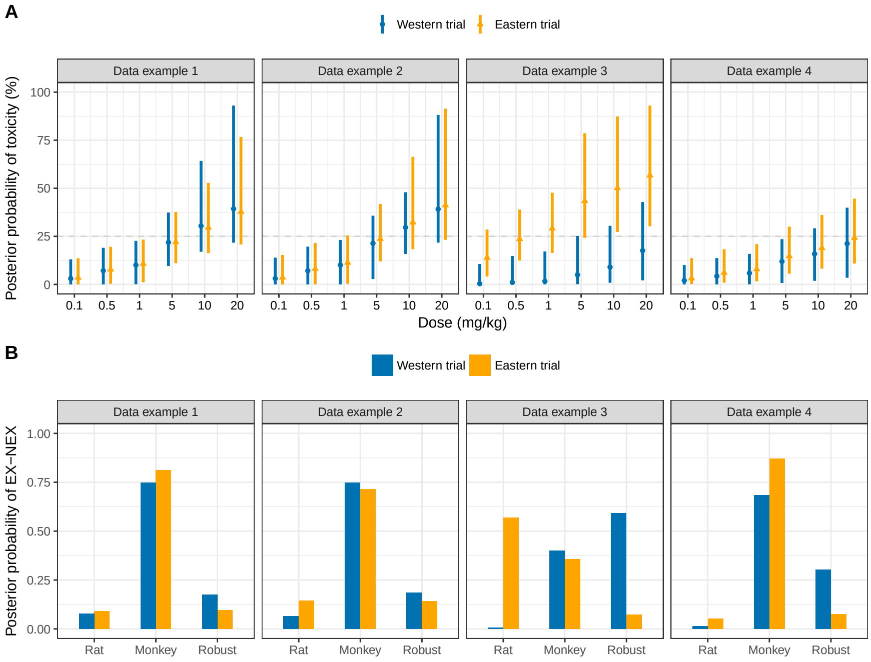

Figure 4.7: Summaries about the posteriors for probability of toxicity synthesising preclinical animal data and human toxicity data (Panel A) and posterior probabilities of exchangeability or non-exchangeability by the end of the hypothetical global trials (Panel B).

posterior probabilities allocated to the EX distribution relating to rat data have been updated as 0.013 and 0.052, respectively. In this scenario, between-trial heterogeneity was acknowledged by compromising between the EX distribution relating to monkey data and the NEX distribution. At the end, dose 20 mg/kg was correctly estimated as the MTD for both trials. We note that incoherent escalation and de-escalation of doses occur in the eastern trial of data examples 4 and 3, respectively. This was due to the relatively informative prior distributions placed on the variance parameters τ_3 and τ_4, which resulted in too much sharing of information between the two phase I trials. The dose-escalation procedure can be improved to be coherent by adopting less weakly informative priors for τ_3 and τ_4, or choosing larger prior probability of NEX for each phase I trial.

4.4 SIMULATION STUDY

In this section, we compare operating characteristics of the global phase I dose-escalation trials, conducted and analysed using either the proposed methodology

Table 4.2: Simulation scenarios for the true probability of toxicity in humans set for the phase I trials $\mathcal{R}_L$ and $\mathcal{R}_B$.

	Western trial, $\mathcal{R}_L$						Eastern trial, $\mathcal{R}_B$					
	d_{L1}	d_{L2}	d_{L3}	d_{L4}	d_{L5}	d_{L6}	d_{B1}	d_{B2}	d_{B3}	d_{B4}	d_{B5}	d_{B6}
	0.1	0.5	1	5	10	20	0.1	0.5	1	5	10	20
Scenario 1	0.01	0.03	0.10	**0.25**	0.34	0.47	0.01	0.03	0.10	**0.25**	0.34	0.47
Scenario 2	0.01	0.03	0.10	**0.25**	0.34	0.47	0.05	0.12	**0.25**	0.37	0.50	0.60
Scenario 3	0.01	0.03	0.10	**0.25**	0.34	0.47	0.01	0.03	0.07	0.15	**0.25**	0.37
Scenario 4	0.01	0.03	0.05	0.08	0.15	**0.25**	0.02	0.05	0.07	0.12	**0.25**	0.36
Scenario 5	**0.25**	0.34	0.47	0.55	0.65	0.75	0.40	0.50	0.60	0.70	0.80	0.90
Scenario 6	0.01	0.03	0.05	0.08	0.15	**0.25**	0.10	**0.25**	0.36	0.50	0.60	0.68

or alternative Bayesian models that may be considered. This simulation study was designed straightly from the illustrative data examples in Section 4.3: we held the animal data and the structure of trials $\mathcal{R}_L$ and $\mathcal{R}_B$ unchanged for all simulated trials. The Bayesian analysis models for comparison are defined as follows.

- Model A: the EX-NEX approach newly proposed in this chapter, where the phase I trials $\mathcal{R}_L$ and $\mathcal{R}_B$, respectively, have their own NEX distribution for addressing possibility of inconsistency between preclinical and clinical data. Here, we stipulate a high level of prior confidence in the available animal data;

- Model B: a standard hierarchical model that assumes full exchangeability of the dose-toxicity parameters $\boldsymbol{\gamma}_\ell$ underpining the phase I trials. Animal data are not incorporated in this model;

- Model C: stratified analysis, where phase I trials $\mathcal{R}_L$ and $\mathcal{R}_B$ are conducted and analysed separately. Animal data are not incorporated in this model;

- Model D: stratified analysis, where phase I trials $\mathcal{R}_L$ and $\mathcal{R}_B$ are conducted and analysed separately. Animal data are leveraged using the Bayesian hierarchical model proposed in Chapter 3 with high level of prior confidence.;

- Model E: Pooling data from trials $\mathcal{R}_L$ and $\mathcal{R}_B$, setting a weakly-informative prior to implemente the dose-escalation procedure. In other words, animal data are discarded completely.

For each iteration of the simulation study, dichotomous toxicity outcomes were generated for each cohort of patients treated with a dose $\hat{d}_L^{(h_{\ell^\star})}$ or $\hat{d}_B^{(h_\ell)}$ in their own region, which fulfils the criterion defined in (4.3.1), from a binary distribution given the true probability of toxicity listed in Table 4.2. Six representative scenarios were considered to characterise the cases of consistency or inconsistency within the patient groups, and that between animal and human toxicity data. The simulated phase I trials may be (i) terminated early for the drug being overly toxic even at the lowest dose, specifically, when, for example $\mathbb{P}(p_{\ell 1} \geqslant 0.33|\mathbf{Y}_1, \ldots, \mathbf{Y}_5, \mathbf{Y}_L^{(h_{\ell\star}-1)}, \mathbf{Y}_B^{(h_\ell-1)}) > 0.25$ at any interim analysis, or (ii) ended after 24 patients were treated with a declaration of MTD for each trial. These two subsets of trials will later be referred to as *stopped early* and *completed* trials, respectively. We base the inference on the region-specific MTD on the point estimate (posterior median) of the posterior distribution for the DLT risk on any dose $d_{\ell j} \in \mathcal{D}_\ell$, denoted by $\tilde{p}_{\ell j}$, at the end of a *completed* trial ℓ:

$$\hat{d}_{\ell M} = \arg \min_{d_{\ell j} \in \mathcal{D}'_\ell} |\tilde{p}_{\ell j} - 0.25|,$$

where $\mathcal{D}'_\ell \subseteq \mathcal{D}_\ell$ contains all the doses that have been given to treat patients and meanwhile satisfy the defined overdose criterion by the end of a *completed* trial ℓ.

For each toxicity scenario in humans and Bayesian analysis model, we simulated 2 000 adaptive phase I dose-escalation trials in each geographic region. Results were reported by summarising across the simulated trials in each geographic region in terms of percentage of trials that were *stopped early* for safety, and percentage of trials that claimed on each dose in the dosing set as MTD. Moreover, we also report the average number of patients allocated to each dose in trials $\mathcal{R}_L$ and $\mathcal{R}_B$, respectively, across the 2 000 simulated trials; that is, both the *completed* and *stopped early* trials were counted. This makes sense as investigators need to know how many patients would be involved in an excessively toxic scenario such as Scenario 5, especially for the eastern trial $\mathcal{R}_B$.

To elucidate the strength and weakness of the proposed methodology for leveraging co-data, we particularly focus on the comparison of performance of Models A – C in the main manuscript. Nevertheless, complete numerical results of this simulation study are listed in Table 4.3 of the Supplementary Materials. Figure 4.8 presents the operating characteristics of phase I dose-escalation trials designed based on Models A – C, respectively. We see that Model A in general is the winner across nearly all the simulation scenarios: it performs well in percentage of correct selection (PCS) of the region-specific MTDs and allocates most patients in a trial to dose(s) associated with the DLT risk(s) falling into the target interval [0.16, 0.33). In Scenario 5 where the drug is overly toxic in patients of the eastern trial, Model A yielded 33.1% and 51.3% of the 2 000 simulated trials in each region to *stop early*, and another 36.7% to be completed with a correct declaration of MTD for the western trials. Scenarios 4 and 6 are more demanding in trial sample size, compared with the rest, as the true MTD

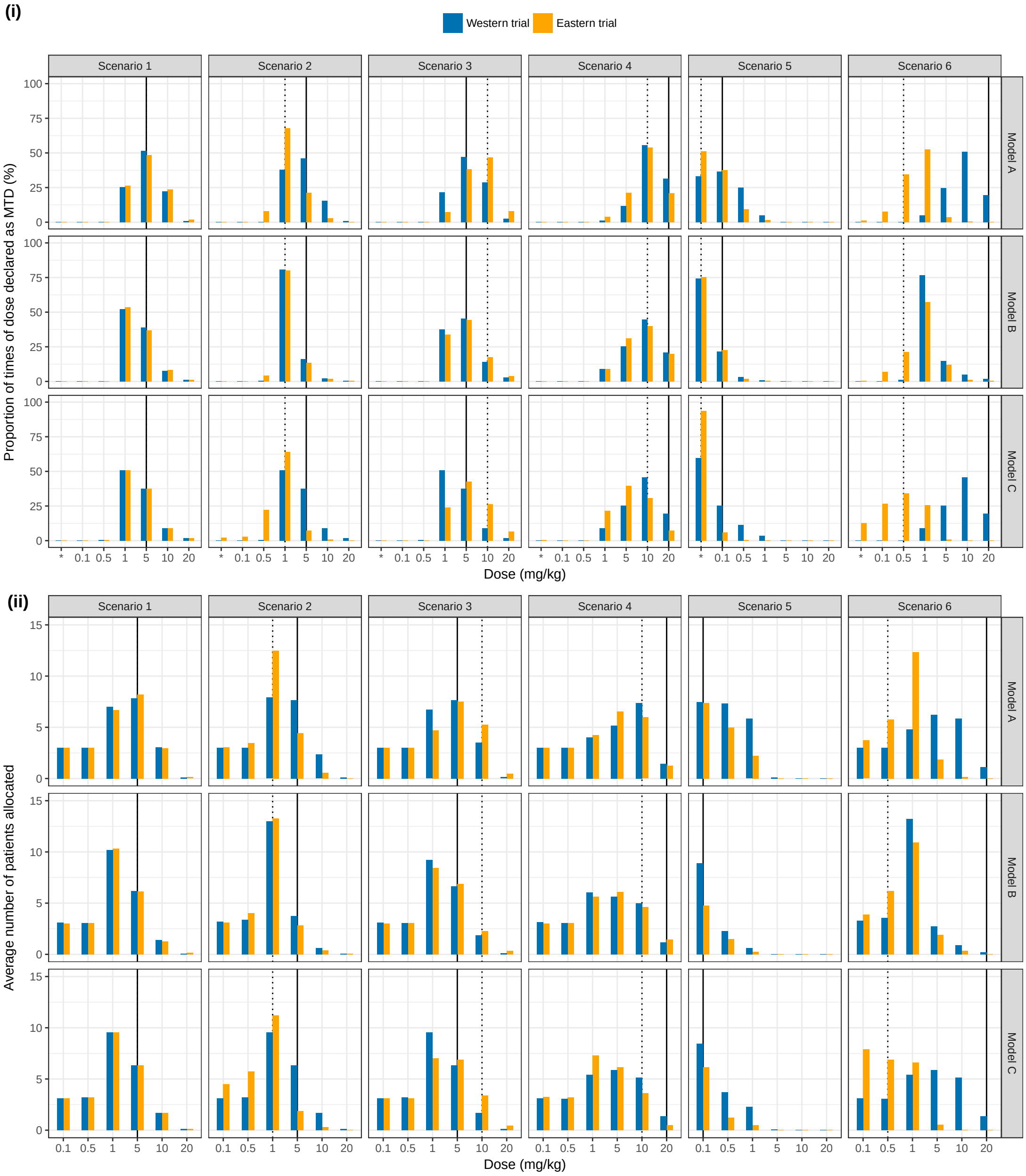

Figure 4.8: Operating characteristics of the adaptive phase I dose-escalation trials in two geographic regions, conducted and analysed using Models A – C. The vertical black solid (dotted) line indicates the true MTD in the western trial $\mathcal{R}_L$ (eastern trial $\mathcal{R}_B$) under each simulation scenario.

for the western trial is the highest dose in set $\mathcal{D}_\ell$. With the "no-skipping-of-dose" restriction, more patients will be needed so that dose for administration can escalate up to 20 mg/kg. It is then not surprising that dose 10 mg/kg was more often claimed to be the MTD for western trials in these two simulation scenarios, as we can read very few patients have been treated with dose 20 mg/kg.

Comparing Model A with Model B in Scenarios 1 – 4 where bridging assumption is correct, benefit from leveraging preclinical animal data and enabling possibility of non-exchangeability is obvious. In Scenario 1, more patients were treated with dose 1 mg/kg than the true MTD 5 mg/kg, due to conservative rules adopted for dose escalations and for concluding on the MTD. Consequently, more trials were ended with a safer dose to be declared as the MTD. In Scenarios 2 – 4, Model B experienced difficulty to distinguish the region-specific MTDs when difference is small. Indeed, it led to excessive sharing of information between the two phase I clinical trials. Model A, in contrast, has reacted sensibly that as a consequent the PCS in both regions was significantly increased. In Scenario 5, more trials designed using Model B were *stopped early* for safety. These two analysis models gave very divergent results in Scenario 6: Model A spot the evident difference in toxicity of the drug to patients in different regions, while Model B underestimated the degree of heterogeneity and ended up with concluding on a dose in the middle ground.

Comparing Model A with Model C, we perceive how much merit is to be attributed to (i) bridging strategies and/or (ii) sharing of information between animals and humans. Under Model C, the PCS for Scenario 1 was 37.5% for simulated trials in each region. This figure increased to 51.6% and 48.4%, respectively, if Model A had been used. In Scenario 2, the gains were mainly from leveraging consistent animal data to facilitate estimating the MTDs. When the drug is overly toxic to patients, Model C tend to be more cautious then Model A, which incorporated animal data suggesting the drug to be safer than it actually was in this simulation study. In the rest of the simulation scenarios, we observed that Model A outperformed Model C for making use of external information.

Results generated from phase I adaptive trials designed using Model D or Model E are given in the Supplementary Materials. Readers may compare Models A and D to see gains from using bridging strategies, and compare Models C and D to see whether leveraging animal data would help estimate dose-toxicity parameters of human trials. When using Model E, estimating region-specific MTDs is not possible, as this is a one-size-fits-all solution. Instead, a single MTD will be concluded. This approach is certainly inappropriate especially in cases of Scenario 6.

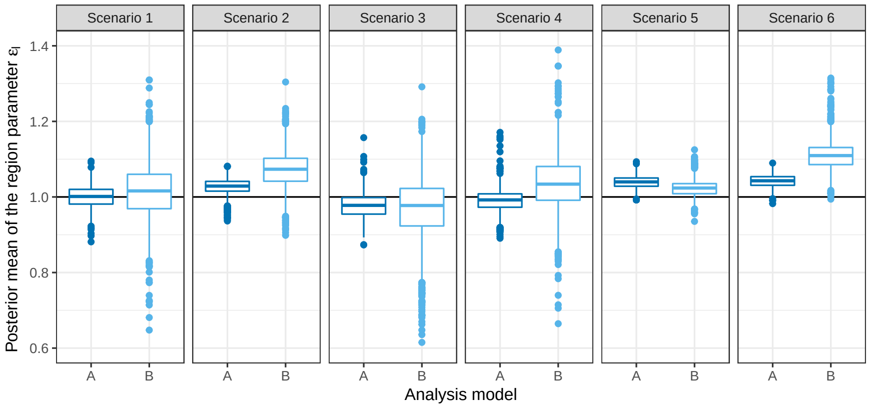

Figure 4.9: Boxplots that depict the posterior means of the region parameter ϵ_ℓ estimated by the end of *completed* trials, designed using Model A or Model B. The horizontal black line represents the prior mean of ϵ_ℓ.

We have introduced a region parameter ϵ_ℓ into the logistic dose-toxicity model to take account of *intrinsic* ethnic differences between patient groups. It would be interesting to estimate this parameter by the end of a *completed* trial. Figure 4.9 shows the boxplots depicting the point estimate (posterior mean) of ϵ_ℓ obtained from the simulated trials have completed treatment for 24 patients in each region. These posterior means are to be compared with the prior mean marked with a horizontal black line that a large difference challenges the appropriateness of using bridging strategies in the global phase I clinical trials. For example, in Scenario 6, when assuming full exchangeability of model parameters γ_ℓ (under Model B), most posterior means took a value greater than the prior mean. This indicates that the drug appears to be more toxic in patients of the eastern trial than those of the western trial. Within the same scenario, same interpretation works for the posterior means of ϵ_ℓ obtained under Model A.

Across all the simulation scenarios except Scenario 5, the size of the shift from the prior to posterior under Model A is smaller than Model B, as inconsistency can also be addressed by the non-exchangeability distribution following specification of Model A. In Scenario 5, more trials were *completed* with correct estimation of the region-specific MTDs under Model A than under B; posterior means of ϵ_ℓ suggest there was a difference in toxicity of the drug. Wheras, in the same scenario, Model B yielded 74.4% and 75.4% of the simulated western and eastern trials to *stop early* for safety, respectively, leaving other 21.8% and 22.5% to proceed until reaching the maximum planned sample size. Model B failed to spot the difference between the region-specific MTDs but concluded on an identical dose by the end of those *com-*

pleted trials. Very small update from prior mean to posterior mean of ϵ_ℓ in Scenario 5 has also suggested the drawback of Model B that it implements excessive sharing of information, as we interpreted earlier. In Scenario 1, no substantial difference between prior and posterior means of ϵ_ℓ was observed, reflecting the correctness of the bridging assumption.

Saving sample size may be appealing to investigators especially when the bridging assumption is correct. We have evaluated possibility of our methodology to permit the second trial to stop early for precision of estimators. Average number of patients needed for the second trial could be reduced in consistency scenarios. We have also run simulations where a different level of prior confidence in the animal data for Model A is considered, namely, setting $w_{\ell \mathrm{Rat}} = 0.10$, $w_{\ell \mathrm{Monkey}} = 0.40$ and $w_{\ell \mathrm{R}} = 0.50$. Conclusions are similar with those written in Chapter 3, and thus will not be repeated in the present chapter.

4.5 APPLICATION TO TRIALS WITH A SEQUENTIAL BRIDGING STRATEGY

In the hypothetical data examples and the simulation study, we have supposed the second trial would begin before the termination of the first trial. Numerical results elucidated advantages of our approach compared with alternative analysis models. In this section, we consider application of this robust Bayesian hierarchical model to phase I trials with a sequential bridging strategy. Specifically, design and analysis of the first and the second trials are concerned in a row.

Suppose that a total of 24 patients will be recruited and treated in a sequence of eight cohorts in the western trial $\mathcal{R}_\mathrm{L}$. From the nineth global cohort, of which the regional cohort number $h_\ell = 1$ if coded locally, and onwards is the timeline for the eastern trial $\mathcal{R}_\mathrm{B}$. Throughout the clinical drug development program, complete information from animal studies (same as what was used in Sections 3 and 4) has been made available. The dose escalation criterion at the interims is then updated as:

$$\begin{aligned} \hat{d}_\mathrm{L}^{(h_{\ell\star})} &= \max\{d_{\ell j} \in \mathcal{D}_\ell : \mathbb{P}(p_{\ell j} \geqslant 0.33 | \mathbf{Y}_1, \ldots, \mathbf{Y}_5, \mathbf{Y}_\mathrm{L}^{(h_{\ell\star}-1)}) \leqslant 0.25\}, \\ \hat{d}_\mathrm{B}^{(h_\ell)} &= \max\{d_{\ell j} \in \mathcal{D}_\ell : \mathbb{P}(p_{\ell j} \geqslant 0.33 | \mathbf{Y}_1, \ldots, \mathbf{Y}_5, \mathbf{Y}_\mathrm{L}, \mathbf{Y}_\mathrm{B}^{(h_\ell-1)}) \leqslant 0.25\}. \end{aligned} \tag{4.5.1}$$

Stipulating same priors for the parameters needed to implement the model, we generate four hypothetical trial examples following a sequential strategy. Particularly, these new trial examples were simulated from same specification of true toxicities in humans as that was considered in Figure 4.6 of Chapter 4.3.2.

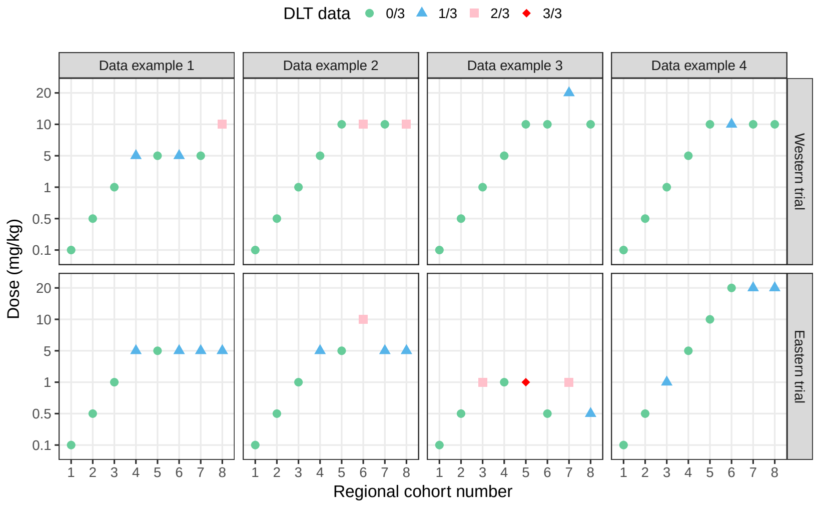

Figure 4.10: Dose-escalation scheme in each phase I clinical trial designed sequentially using the proposed Bayesian EX-NEX model.

Figure 4.10 gives an overview of how these trials may progress. Here, we have forced both trials to start with a safe dose 0.1 mg/kg and not permitted escalation with skipping of a dose. We observe that the trial trajectories were not much different from what was obtained under a parallel bridging strategy adopted in previous sections of this paper. This suggest inferences based on our methodology are robust to the toxicity data of current trial. During the course of the western trial $\mathcal{R}_L$, no dose-toxicity data were generated from the eastern trial $\mathcal{R}_B$ for the period of global cohort 1 – 8. There were thus no information to update priors for the heterogeneity between human trials. But this does not preclude implementation of our model, as we have used proper priors that integrates to 1 for probablistic inferences. We would like to use these additional data examples to generalise the conclusions obtained from previous numerical evaluations where the trial setting appears to be a bit restrictive.

4.6 DISCUSSION

Bridging studies have been widely discussed in the statistical literature (Wadsworth et al. [2018]), as they show promise to demonstrate drug behaviours using fewer resources rather than establishing an independent package of clinical drug development. To date, much methodology work to extrapolate foreign clinical data to the population of a new region has focused on settings of phase II and phase III trials;

see, for example, Hsiao et al. [2004]; Chow et al. [2012]; Tsou et al. [2012]; Zhang et al. [2017]. Discussion on this topic in the context of phase I dose-escalation trials was limited. In this chapter, we have presented a flexible Bayesian EX-NEX approach for leveraging co-data available from preclinical animal studies and phase I clinical trials, either completed or ongoing, in a different region. For illustration, we showed numerical studies considering two phase I trials to be undertaken in distinct geographic regions, while the methodology can certainly be used for design and analysis of more phase I clinical trials to estimate the region-specific MTDs.

When evaluating properties of our Bayesian EX-NEX approach in Chapters 4.3 and 4.4, we have assumed same recruitement rate for all trials involved in the global program. However, this is not a requirement to implement the model, as potentially heterogeneous preclinical and clinical data are synthesised through the K EX distributions for model parameters (rather than the data), stipulated in the top hierarchies. We also note current version of our model is not readily to give accurate estimates of population-averaged effects in humans, as we did not explicitly parameterise any population means specific to humans for summarising findings of several trials conducted in the same region.

The methodology proposed in this work may also be applied in exploratory phase basket trials (Thall et al. [2003]; Berry et al. [2013]). Historical data would need to be carefully selected to formulate the EX distributions for the parameters underpinning each basket, where the co-data are to be generated. We imagine that leveraging co-data on different endpoints across baskets would increase complexity. While our Bayesian EX-NEX approach is limited in this aspect, further research to extend this model to accommodate correlated but not identical endpoints could be of interest.

4.7 TECHNICAL NOTES

4.7.1 *OpenBUGS code for implementation*

```
model{
# likelihood/sampling model
# MdoseA: total number of doses tested in animal species
for(j in 1:MdoseA){
linA[j] <- theta[StudyA[j], 1]
      + exp(theta[StudyA[j], 2])*log(deltaA[Species[j]]*DoseA[j]/DoseRef)
```

```
logit(pToxA[j]) <- linA[j]
NtoxA[j] ~ dbin(pToxA[j], NsubA[j])
}

zero[1] <- 0
zero[2] <- 0

# theta=(theta1, theta2) derived from each animal study are ready for the use
# on the human equivalent scale
for(i in 1:NstudyA){
for(j in 1:MdoseH){
lin[i, j] <- theta[i, 1] + exp(theta[i, 2])*log(DoseH[j]/DoseRef)
}

# sp.ind[i]: index function to specify
# which species the Study i belongs to
theta[i, 1] <- mu.sp[sp.ind[i], 1] + re.A[i, 1]
theta[i, 2] <- mu.sp[sp.ind[i], 2] + re.A[i, 2]
re.A[i, 1:2] ~ dmnorm(zero[1:2], prec.Psi[1:2, 1:2])

# PInd[]: matrice of the trivial/non-trivial weights
# trivial weights for animals, no local robustification
# to assure theta_i are fully exchangeable within the same species
sp.ind[i] ~ dcat(PInd[i, 1:(n.sp+1)])
}

# Animal species cluster
for(k in 1:n.sp){
deltaA[k] <- exp(Prior.mn.deltaA[k]
                  + Prior.sd.deltaA[k]*log.deltaA01[k])
log.deltaA01[k] ~ dnorm(0, 1)
mu.sp[k, 1] <- mu[1] + re.m[k, 1]
mu.sp[k, 2] <- mu[2] + re.m[k, 2]
re.m[k, 1:2] ~ dmnorm(zero[1:2], prec.Sigma[1:2, 1:2])

theta.predH[k, 1] <- mu.sp[k, 1] + re.h[k, 1]
theta.predH[k, 2] <- mu.sp[k, 2] + re.h[k, 2]
re.h[k, 1:2] ~ dmnorm(zero[1:2], prec.Phi[1:2, 1:2])
}

# default weakly-informative prior for robustification
theta.predH[(n.sp+1), 1:2] ~ dmnorm(Prior.mw[1:2], prec.sw[1:2, 1:2])
cov.rb[1, 1] <- pow(Prior.sw[1], 2)
cov.rb[2, 2] <- pow(Prior.sw[2], 2)
cov.rb[1, 2] <- Prior.sw[1]*Prior.sw[2]*Prior.corr
cov.rb[2, 1] <- cov.rb[1, 2]
prec.sw[1:2, 1:2] <- inverse(cov.rb[1:2, 1:2])

# Meta-analytic prediction
for(i in 1:n.sb){
theta.star[i, 1] <- theta.predH[exch.index[i], 1]
theta.star[i, 2] <- theta.predH[exch.index[i], 2]

# latent mixture indicators:
# exch.index: categorical 1, ..., (n.sp+1)
exch.index[i] ~ dcat(wMix[1:(n.sp+1)])
for(ii in 1:(n.sp+1)){
each[i, ii] <- equals(exch.index[i], ii)
}

# ADD HUMAN DATA HERE
for(j in 1:MdoseH){
linH[i, j] <- theta.star[i, 1]
                + exp(theta.star[i, 2])*log(deltaH[i]*DoseH[j]/DoseRef)
logit(pToxH[i, j]) <- linH[i, j]
NtoxH[i, j] ~ dbin(pToxH[i, j], NsubH[i, j])

pCat[i, j, 1] <- step(pTox.cut[1] - pToxH[i, j])
pCat[i, j, 2] <- step(pTox.cut[2] - pToxH[i, j])
               - step(pTox.cut[1] - pToxH[i, j])
pCat[i, j, 3] <- step(1 - pToxH[i, j])
               - step(pTox.cut[2] - pToxH[i, j])
}

deltaH[i] <- Prior.mn.deltaH[i] + Prior.sd.deltaH[i]*deltaH01[i]
deltaH01[i] ~ dnorm(0, 1)I(-3.921, 3.921)
}

# priors: Prior.mt1, Prior.mt2
prec.mt1 <- pow(Prior.mt1[2], -2)
prec.mt2 <- pow(Prior.mt2[2], -2)

# numerical stability:
# constrained to -10 and +10 (mt1), -5 and +5 (mt2)
mu[1] ~ dnorm(Prior.mt1[1], prec.mt1)I(-10, 10)
mu[2] ~ dnorm(Prior.mt2[1], prec.mt2)I(-5, 5)

# Priors for hyper parameters of the covariance matrix
# say, prec.Psi[1:2, 1:2]
prec.tau1 <- pow(Prior.tau.HN[1], -2)
prec.tau2 <- pow(Prior.tau.HN[2], -2)
tauA[1] ~ dnorm(0, prec.tau1)I(0.001,)
tauA[2] ~ dnorm(0, prec.tau2)I(0.001,)

covA.ex[1, 1] <- pow(tauA[1], 2)
covA.ex[2, 2] <- pow(tauA[2], 2)
covA.ex[1, 2] <- tauA[1]*tauA[2]*rhoA
covA.ex[2, 1] <- covA.ex[1, 2]
prec.Psi[1:2, 1:2] <- inverse(covA.ex[1:2, 1:2])

rhoA ~ dunif(Prior.rho[1], Prior.rho[2])

# Priors for hyper parameters of the covariance matrix
# say, prec.Sigma[1:2, 1:2]
prec.sigma1 <- pow(Prior.sigma.HN[1], -2)
prec.sigma2 <- pow(Prior.sigma.HN[2], -2)
sigma[1] ~ dnorm(0, prec.sigma1)I(0.001,)
sigma[2] ~ dnorm(0, prec.sigma2)I(0.001,)
```

```
covA.sig[1, 1] <- pow(sigma[1], 2)
covA.sig[2, 2] <- pow(sigma[2], 2)
covA.sig[1, 2] <- sigma[1]*sigma[2]*kappaA
covA.sig[2, 1] <- covA.sig[1, 2]
prec.Sigma[1:2, 1:2] <- inverse(covA.sig[1:2, 1:2])

kappaA ~ dunif(Prior.kappa[1], Prior.kappa[2])

# Priors for hyper parameters of the covariance matrix
# say, prec.Phi[1:2, 1:2]
prec.tau3 <- pow(Prior.tau.HN[3], -2)
prec.tau4 <- pow(Prior.tau.HN[4], -2)
tauH[1] ~ dnorm(0, prec.tau3)I(0.001,)
tauH[2] ~ dnorm(0, prec.tau4)I(0.001,)

covH.ex[1, 1] <- pow(tauH[1]], 2)
covH.ex[2, 2] <- pow(tauH[2], 2)
covH.ex[1, 2] <- tauH[1]*tauH[2]*rhoH
covH.ex[2, 1] <- covH.ex[1, 2]
prec.Phi[1:2, 1:2] <- inverse(covH.ex[1:2, 1:2])

rhoH ~ dunif(Prior.rho[1], Prior.rho[2])
}
```

4.8 SUPPLEMENTARY MATERIALS

4.8.1 *Graphical representation of the toxicity scenarios investigated in humans*

4.8.2 *Additional simulation results*

Leveraging co-data, which are consistent with current trial data, should improve the accuracy of posterior estimates for the probability of toxicity, compared with analysis without borrowing of information at all. On the other hand, when co-data would be inconsistent with current trial data, we are certainly much concerned with the performance of posterior estimates. To this end, we obtain the point estimates (poste-

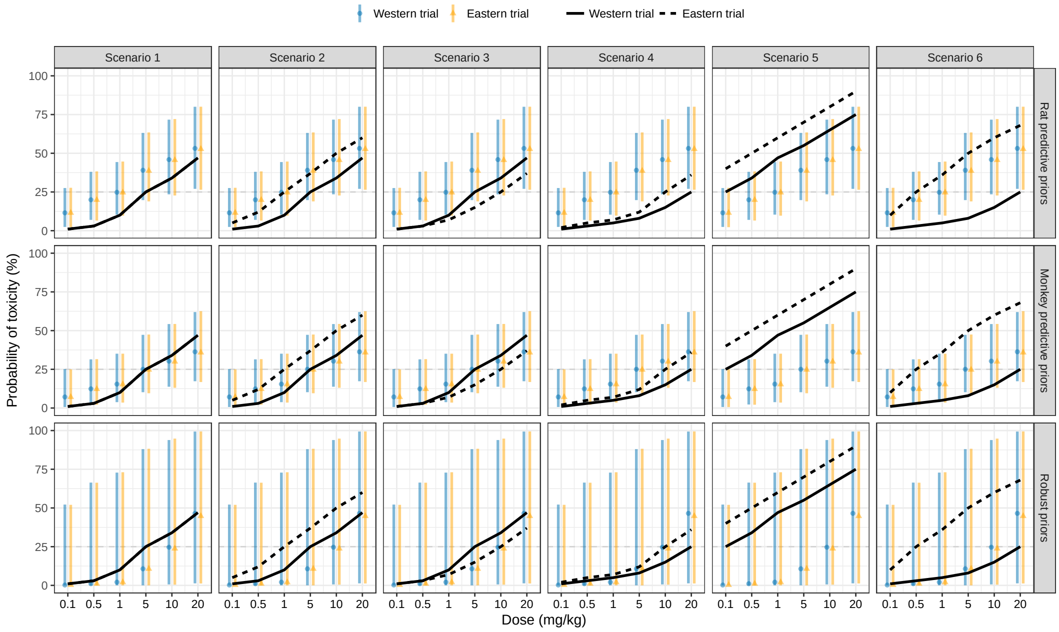

Figure 4.11: Toxicity scenarios in humans that have been considered in the simulation study, overlaying the predictive priors obtained from the animal studies or weakly informative priors.

rior medians) by the end of the *completed* trials, which constitute one subset of the 2 000 simulated trials in each region, and compute the arithmetic means of such point estimates per human dose under the analysis models A – E.

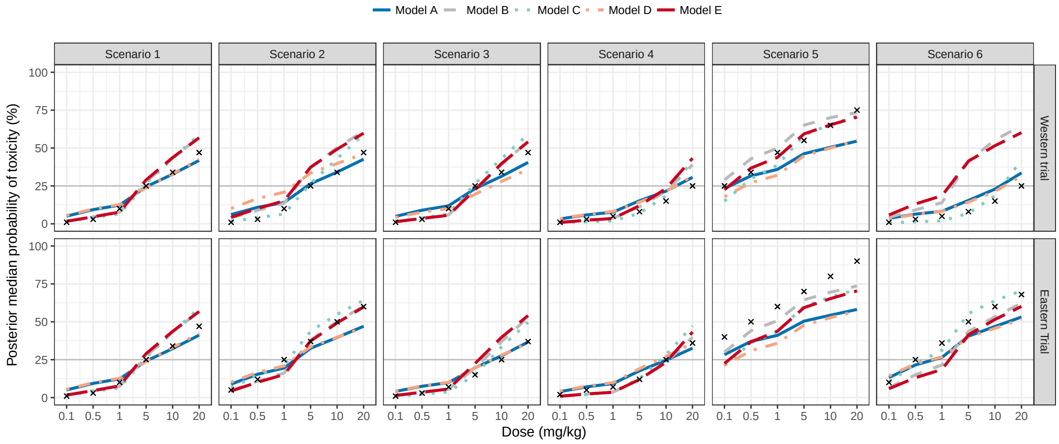

Figure 4.12: Average fitted dose-toxicity curves obtained under Models A – E, based on the *completed* trials only. The black cross marks the true probability of toxicity per human dose of interest for the six simulation scenarios. The horizontal gray line indicates the target level, at which the MTD is defined.

Figure 4.12 visualises how close the estimated dose-toxicity relationships (obtained by averaging across the *completed* trials) would be, to the true probability of toxicicty in humans marked with black cross per dose of interest. Since it was not possible to estimate region-specific MTDs under Model E which pools western and eastern data together for analysis, the average fitted curves in red for Model E displayed in two regions are identical within same simulation scenario.

As we can see, Model B and Model E have presented less satisfactory performance, compared with the rest, across nearly all the simulation scenarios. The dose-toxicity curves fitted using these two models are severely deviated from the true dose-toxicity relationship when the region-specific MTDs are very divergent, such as in Scenario 6. Model B bears excessive sharing of information, which as a result is inappropriate for the divergent scenarios. In contrast, Models A, C – D give quite robust estimates about the dose-toxicity relationship in most cases. Moreover, these models converge well to the true MTD. Scenario 5 corresponds to an overly toxic situation, where most simulated trials are meant to be *stopped early*. The complement subset, i.e., *completed* trials, are those suggested the drug to be less toxic given fewer DLTs observed. Consequently, the fitted curves obtained from the *completed* trials only tend to underestimate the true probability of toxicity.

In the main mauscript, we have interpreted gains from Model A over Model C as a mixture of advantages to the use of both animal data and bridging strategy. Here, we add Model D into the comparison to see how much gains should be attributed to using animal data (comparing Models C and D), and the robust bridging strategy (comparing Models A and D). Figure 4.13 show operating characteristics of the global phase I dose-escalation trials, planned using Models A, C – D. Numerical results of all the analysis models are listed in Tabel 4.3, where we have also included the non-parametrical optimal benchmark design (O'Quigley [2002]) for comparison. Across all the analysis models, our proposed Model A presents very similar behaviour in PCS with that of the benchmark design in consistency scenarios. In Scenario 2, where animal data are very consistent with human data and the bridgin assumption holds, we see that the dose-escalation design based on Model A defeats the benchmark design in PCS for the eastern trial. In contrast, referring to the results in Scenario 6, enabling incorporation of co-data from external studies does not help much in identifying the true MTD. This leads to a larger deviation between properties of the dose-escalation design based on Model A and the optimal benchmark design.

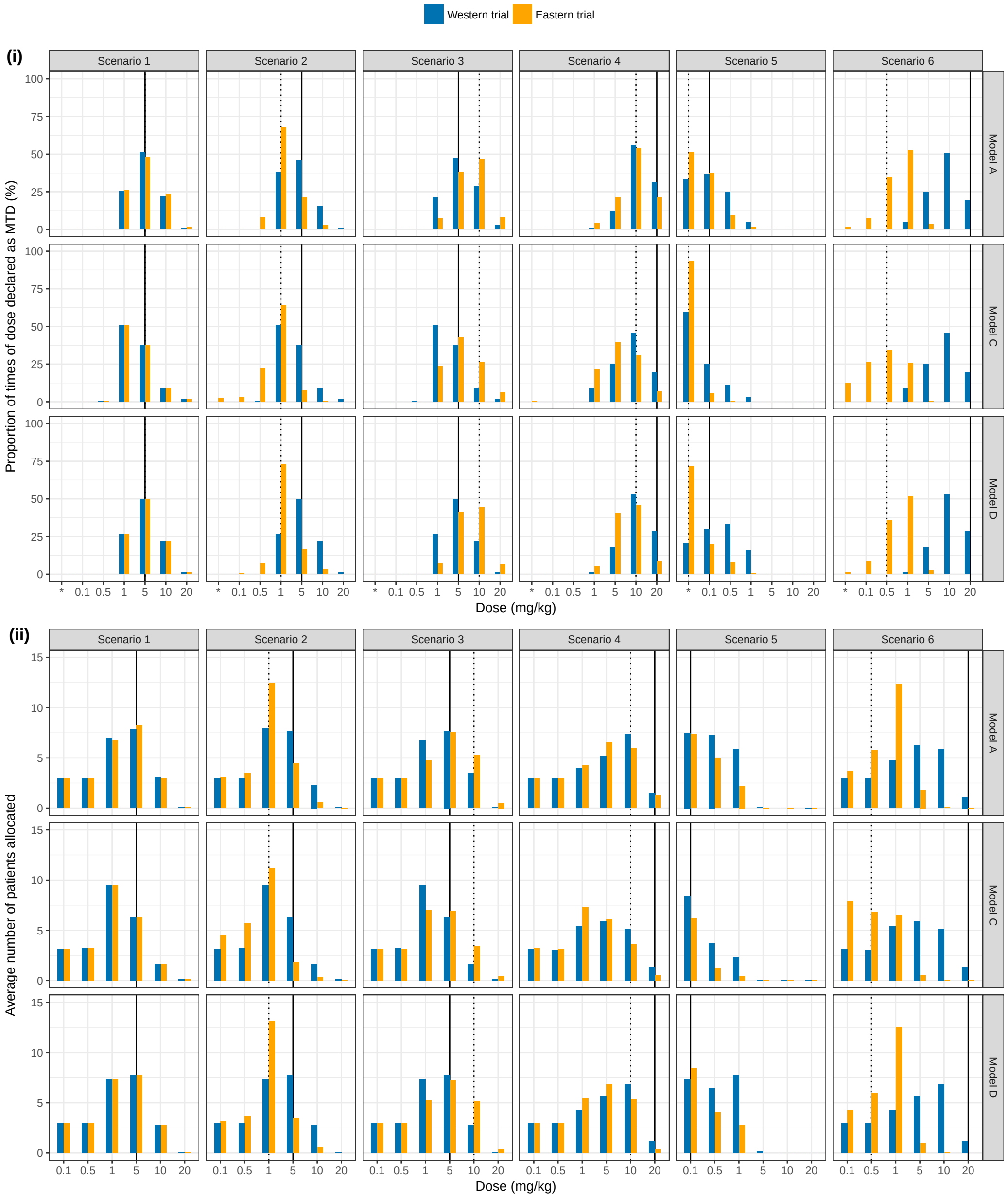

Figure 4.13: Operating characteristics of the adaptive phase I dose-escalation trials in two geographic regions, conducted and analysed using Models A, C, D. The vertical black solid (dotted) line indicates the true MTD in the western trial $\mathcal{R}_L$ (eastern trial $\mathcal{R}_B$) under each simulation scenario.

Table 4.3: Comparison of alternative analysis models in terms of the percentage of selecting a dose as MTD at the end of the trials, percentage of early stopping for safety, average patient allocation, and average number of patients with toxicity.

Design			% dose declared as MTD & average patient allocation															
			Western trial, $\mathcal{R}_L$							Eastern trial, $\mathcal{R}_B$							$\bar{N}_L$	$\bar{N}_B$
			d_{L1} 0.1	d_{L2} 0.5	d_{L3} 1	d_{L4} 5	d_{L5} 10	d_{L6} 20	None	d_{B1} 0.1	d_{B2} 0.5	d_{B3} 1	d_{B4} 5	d_{B5} 10	d_{B6} 20	None		
Scenario 1		pTox	0.01	0.03	0.10	**0.25**	0.34	0.47		0.01	0.03	0.10	**0.25**	0.34	0.47			
	Benchmark	Sel	0	0.3	14.6	**59.1**	22.0	4.0		0	0.3	14.6	**59.1**	22.0	4.0			
	Model A	Sel	0	0	25.2	**51.7**	22.3	0.9	0	0	0	26.4	**48.4**	23.5	1.7	0		
		Pts	3.0	3.0	7.0	**7.9**	3	0.1		3.0	3.0	6.7	**8.2**	3.0	0.1		24.0	24.0
	Model B	Sel	0	0	52.2	**39.0**	7.7	1.1	0	0	0	53.5	**37.1**	8.3	1.1	0		
		Pts	3.0	3.1	10.2	**6.2**	1.4	0.1		3.0	3.1	10.3	**6.2**	1.3	0.1		24.0	24.0
	Model C	Sel	0	0.6	50.8	**37.5**	9.1	1.8	0.2	0	0.6	50.8	**37.5**	9.1	1.8	0.2		
		Pts	3.1	3.2	9.5	**6.3**	1.7	0.1		3.1	3.2	9.5	**6.3**	1.7	0.1		23.9	23.9
	Model D	Sel	0	0	26.8	**50.0**	22.1	1.1	0	0	0	26.8	**50.0**	22.1	1.1	0		
		Pts	3.0	3.0	7.4	**7.7**	2.8	0.1		3.0	3.0	7.4	**7.7**	2.8	0.1		24.0	24.0
	Model E	Sel	0	0	59.7	**33.1**	5.9	1.1	0	-	-	-	-	-	-	-		
		Pts	6.0	6.2	21.8	**11.8**	1.9	0.2		-	-	-	-	-	-		47.9	-
Scenario 2		pTox	0.01	0.03	0.10	**0.25**	0.34	0.47		0.05	0.12	**0.25**	0.37	0.50	0.60			
	Benchmark	Sel	0	0.3	14.6	**59.1**	22.0	4.0		0.7	18.9	**56.8**	21.4	2.2	0			
	Model A	Sel	0	0	37.8	**46.0**	15.4	0.8	0	0.2	7.9	**67.9**	21.1	2.8	0.1	0		
		Pts	3.0	3.0	7.9	**7.7**	2.3	0.1		3.0	3.5	**12.5**	4.4	0.6	0		24.0	24.0
	Model B	Sel	0	0.5	80.7	**16.2**	2.1	0.5	0	0.2	4.2	**80.0**	13.4	1.8	0.4	0		
		Pts	3.2	3.4	13.0	**3.7**	0.6	0.1		3.1	4.1	**13.4**	2.9	0.4	0.1		24.0	24.0
	Model C	Sel	0	0.6	50.8	**37.5**	9.1	1.8	0.2	3.0	22.4	**63.9**	7.4	0.8	0.1	2.4		
		Pts	3.1	3.2	9.5	6.3	1.7	0.1		4.5	5.7	11.2	1.8	0.3	0		23.9	23.5
	Model D	Sel	0	0	26.8	**50.0**	22.1	1.1	0	0.5	7.3	**72.9**	16.2	3.0	0	0.1		
		Pts	3.0	3.0	7.4	7.7	2.8	0.1		3.2	3.6	13.2	3.5	0.5	0		24.0	24.0
	Model E	Sel	0	1.1	**84.6**	**12.0**	0.9	0.3	1.1	-	-	-	-	-	-	-		
		Pts	6.1	7.4	**27.7**	**5.4**	0.7	0.1		-	-	-	-	-	-		47.4	-
Scenario 3			0.01	0.03	0.10	**0.25**	0.34	0.47		0.01	0.03	0.07	0.15	**0.25**	0.37			
	Benchmark	Sel	0	0.3	14.6	**59.1**	22.0	4.0		0	0	2.1	26.7	**48.0**	23.1			
	Model A	Sel	0	0	21.6	**47.2**	28.6	2.6	0	0	0	7.1	38.4	**46.6**	7.9	0		
		Pts	3.0	3.0	6.7	**7.7**	3.5	0.1		3.0	3.0	4.7	7.5	**5.3**	0.5		24.0	24.0

Table 4.3 – *Continued.*

Design		% dose declared as MTD & average patient allocation														$\tilde{N}_L$	$\tilde{N}_B$
		Western trial, $\mathcal{R}_L$							Eastern trial, $\mathcal{R}_B$								
		d_{L1} 0.1	d_{L2} 0.5	d_{L3} 1	d_{L4} 5	d_{L5} 10	d_{L6} 20	None	d_{B1} 0.1	d_{B2} 0.5	d_{B3} 1	d_{B4} 5	d_{B5} 10	d_{B6} 20	None		
Model B	Sel	0	0	37.6	**45.4**	14.1	2.9	0	0	0	34.0	44.5	**17.5**	4.0	0		
	Pts	3.1	3.1	9.2	**6.6**	1.9	0.1		3.0	3.1	8.4	6.9	**2.3**	0.3		24.0	24.0
Model C	Sel	0	0.6	50.8	**37.5**	9.1	1.8	0.2	0	0.1	24.1	42.9	**26.2**	6.5	0.2		
	Pts	3.1	3.2	9.5	**6.3**	1.7	0.1		3.1	3.1	7.0	6.9	**3.4**	0.4		23.9	23.9
Model D	Sel	0	0	26.8	**50.0**	22.1	1.1	0	0	0	7.3	40.9	**44.8**	7.0	0		
	Pts	3.0	3.0	7.4	**7.7**	2.8	0.1		3.0	3.0	5.3	7.2	**5.1**	0.4		24.0	24.0
Model E	Sel	0	0	41.2	**41.4**	**14.3**	2.9	0.2	-	-	-	-	-	-	-		
	Pts	6.0	6.1	18.1	**13.7**	**3.4**	0.6		-	-	-	-	-	-		47.9	-
Scenario 4		0.01	0.03	0.05	0.08	0.15	0.25		0.02	0.05	0.07	0.12	0.25	0.36			
Benchmark	Sel	0.2	0	0.4	3.5	29.3	**66.6**		0	0.5	1.7	17.3	**56.4**	24.1			
Model A	Sel	0	0	1.1	11.8	55.6	**31.5**	0	0	0	4.0	21.2	**53.8**	21.0	0		
	Pts	3.0	3.0	4.0	5.2	7.4	**1.4**		3.0	3.0	4.2	6.5	**6.0**	1.3		24.0	24.0
Model B	Sel	0	0	8.9	25.5	44.7	**20.9**	0	0	0	9.0	31.1	**39.9**	20.0	0		
	Pts	3.1	3.1	6.0	5.6	5	**1.2**		3.1	3.1	5.6	6.1	**4.7**	1.4		24.0	24.0
Model C	Sel	0	0.1	8.9	25.4	45.9	**19.5**	0.2	0.1	0.2	21.6	39.7	**30.6**	7.3	0.4		
	Pts	3.0	3.0	5.4	5.9	5.1	**1.4**		3.2	3.2	7.3	6.1	**3.6**	0.5		23.8	23.9
Model D	Sel	0	0	1.3	17.7	52.6	**28.4**	0	0	0	5.4	40.1	**45.9**	8.6	0		
	Pts	3.0	3.0	4.3	5.7	6.8	**1.2**		3.0	3.0	5.4	6.8	**5.4**	0.4		24.0	24.0
Model E	Sel	0	0	11.9	29.1	**39.1**	**19.6**	0.3	-	-	-	-	-	-	-		
	Pts	6.0	6.2	11.6	12.4	**8.6**	**3.0**		-	-	-	-	-	-		47.8	-
Scenario 5		0.25	0.34	0.47	0.55	0.65	0.75		0.40	0.50	0.60	0.70	0.80	0.90			
Benchmark	Sel	**74.0**	21.9	3.8	0.2	0	0		98.4	1.4	0.2	0	0	0			
Model A	Sel	**36.7**	24.9	5.1	0.2	0	0	33.1	37.7	9.4	1.6	0	0	0	**51.3**		
	Pts	**7.5**	7.3	5.8	0.1	0	0		7.4	5.0	2.2	0	0	0		20.7	14.6
Model B	Sel	**21.8**	3.1	0.8	0	0	0	74.3	22.5	1.8	0.4	0	0	0	**75.3**		
	Pts	**8.9**	2.3	0.6	0	0	0		4.7	1.5	0.2	0	0	0		11.8	6.4
Model C	Sel	**25.2**	11.5	3.5	0	0	0	59.8	5.9	0.4	0.2	0	0	0	**93.5**		
	Pts	**8.4**	3.7	2.3	0.1	0	0		6.2	1.2	0.5	0	0	0		14.5	7.9
Model D	Sel	**29.9**	33.4	16.1	0.2	0	0	20.4	19.9	7.8	0.6	0	0	0	**71.7**		
	Pts	**7.4**	6.4	7.7	0.2	0	0		8.5	4.0	2.8	0	0	0		21.7	15.3

Table 4.3 – *Continued.*

Design			% dose declared as MTD & average patient allocation															$\tilde{N}_L$	$\tilde{N}_B$
			Western trial, $\mathcal{R}_L$							Eastern trial, $\mathcal{R}_B$								
			d_{L1} 0.1	d_{L2} 0.5	d_{L3} 1	d_{L4} 5	d_{L5} 10	d_{L6} 20	None	d_{B1} 0.1	d_{B2} 0.5	d_{B3} 1	d_{B4} 5	d_{B5} 10	d_{B6} 20	None		
	Model E	Sel	**11.1**	0.8	0	0	0	0	**88.1**	-	-	-	-	-	-	-		
		Pts	**11.6**	3.2	0.4	0	0	0		-	-	-	-	-	-		15.2	-
Scenario 6			0.01	0.03	0.05	0.08	0.15	**0.25**		0.10	**0.25**	0.36	0.50	0.60	0.68			
	Benchmark	Sel	0.2	0	0.4	3.5	29.3	**66.6**		14.7	**60.6**	22.2	2.5	0	0			
	Model A	Sel	0	0	5.1	24.6	50.9	**19.4**	0	7.6	**34.6**	52.5	3.5	0.5	0	1.3		
		Pts	3.0	3.0	4.8	6.2	5.9	**1.1**		3.7	**5.8**	12.3	1.8	0.1	0		24.0	23.7
	Model B	Sel	0.1	1.2	76.9	14.8	4.9	**1.9**	0.2	7.0	**21.2**	57.5	12.2	1.3	0.4	0.4		
		Pts	3.3	3.6	13.2	2.8	0.9	**0.2**		3.9	**6.2**	10.9	1.9	0.3	0		24.0	23.2
	Model C	Sel	0	0	8.9	25.4	45.9	**19.5**	0.3	26.7	**34.2**	25.6	0.8	0.1	0	12.6		
		Pts	3.1	3.1	5.4	5.9	5.1	**1.4**		7.9	6.9	6.6	0.5	0	0		24.0	21.9
	Model D	Sel	0	0	1.3	17.7	52.6	**28.4**	0	9.0	**36.1**	51.3	2.4	0	0	1.2		
		Pts	3.0	3.0	4.3	5.7	6.8	**1.2**		4.3	**6.0**	12.5	1.0	0	0		24.0	23.8
	Model E	Sel	0.6	**6.2**	79.3	8.3	1.8	**0.4**	3.4	-	-	-	-	-	-	-		
		Pts	7.0	10.7	25.1	3.2	0.5	0.1		-	-	-	-	-	-		46.6	-

pTox: true probability of toxicity in humans; **Sel**: proportion of times of declaring a dose as MTD; **Pts**: average number of patients allocated to a dose; **Benchmark**: Non-parametrical optimal benchmark design by O'Quigley [2002].

5 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

5.1 SUMMARY OF OUR METHODOLOGIES

Eliciting an informative prior is a widely discussed topic in the Bayesian inference. In this thesis, we have followed the maxim that *Today's posterior is tomorrow's prior*, as Lindley [1972] put it. Focusing on the transition step in early drug development, we have shown how preclinical animal data, seen as a special type of historical data for a new phase I clinical trial, can be used to learn about the toxicity of the same drug in humans. This setting raises several interesting questions. First, how could we translate preclinical animal data that have been recorded on a different dosing scale onto a suitable one to predict toxicity in humans. Second, after careful selection and translation of relevant animal data, how would we cope with the intrinsic differences between toxicity of the drug in animals and humans. Third, what could we do if the predictability of human toxicity varies across animal species, when preclinical data are available from more than one species. Fourth, how could we improve trial efficiency and balance the information from heterogeneous sources when, preceded by preclinical studies, there are more than one phase I dose-escalation trials to be designed.

From Chapter 2 to Chapter 4, we have addressed several facets of these interesting research problems, by developing novel Bayesian adaptive methods to use preclinical animal data in a robust manner during the course of ongoing phase I clinical trials. To the best of our knowledge, our work represents the first applications of Bayesian approaches to leveraging historical data across species, or say, more generally, any historical datasets that have been recorded in very different measurement scales, in human trials. The methodologies relax the requirement which is essential to apply most data augmentation techniques existing in the statistical literature: the source data and the target data for synthesis can be dissimilar with intrinsic and extrinsic variabilities. Our proposals written in different chapters share similarities in that they consist of: (i) informative priors for the toxicity in humans, (ii) a weakly informative prior for the possibility that animal data are not relevant at all, and (iii) a set of prior mixture weights to be allocated to the informative animal priors and the robust prior, respectively. The presented methodologies are, however, quite differential in technical details and how we are concerned with heterogeneity both between studies and between species.

In Chapter 2, we follow the current practice of translational sciences to convert animal doses into equivalent human doses, but formally describe uncertainty that surrounds this translation as well as predictability of the available animal data. The question "Is the drug more toxic in humans than what we have expected from animal studies?" has been asked throughout the phase I first-in-man trial. We have proposed a procedure to assess the commensurability between preclinical and clinical data in a quantifiable manner as the phase I trial progresses. Comparing the prior predictions, obtained based on animal data alone, with the observed toxicity outcomes of patients after treatment, we penalise (reward) incorrect (correct) prior predictions through a small (large) utility. In particular, for the incorrect prior predictions, a smaller utility will be allocated if the accumulating evidence suggests animal data underestimate, rather than overestimate, toxicity in humans. Predictive accuracy of animal data is computed at each human dose for evaluation, and later summarised using an overall quantity to determine the amount of preclinical information to be incorporated. A large prior mixture weight will consequently be allocated to the informative animal prior, comprised in a mixture prior, when the preclinical information gives correct prediction towards human toxicity, and likewise a small prior mixture when it does not. Our Bayesian procedure is shown to be responsive to prior-data conflicts in small trials.

In Chapter 3, we develop a Bayesian meta-analytic approach to incorporate animal data from multiple species into a phase I oncology trial. Since the risk of toxicity could vary drastically across different animal species and humans given the same dose, the exchangeability assumption for model parameters required to establish a hierarchical model would not always hold. We introduce a translation factor into the dose-toxicity models to translate the animal doses onto an equivalent human dosing scale. The model parameters in any tested animal species and humans can then be interpreted on a common scale. Unlike current practice that adopts fixed constants to extrapolate a dose evaluated in animals, we treat the translation factor, appropriate for each animal species, to be a random variable with a log-normal prior to capture uncertainty about the magnitude of such translation across species. Random-effects distributions are stipulated for the parameters, expressed on the common scale, to take account of heterogeneity both between studies and between species. A prior mixture weight is specified representing our prior scepticism about the plausibility of an exchangeability assumption for the human parameters with those estimated from animal studies in a particular species. The possibility of non-exchangeability between human parameters and those of any animal species is also considered. Our methodology is not limited to this particular setting, but can be applied more broadly when the source data have been recorded on a measurement scale different from one where the target data would be generated.

Chapter 4 extends the methodology proposed in Chapter 3 from "M-to-1" settings applied to incorporate preclinical animal data from several studies and species for inference in a phase I first-in-man trial to "M-to-L" settings wherein the decision making pertains to a number of phase I dose-escalation trials. We have illustrated potential applications of our generalised model in the context of phase I clinical trials to be designed and analysed in various geographical regions. A region parameter is introduced to account for intrinsic ethnic differences that could impact on toxicity of a medicine in distinct patient subgroups. We use preclinical animal data to estimate the means of the exchangeability distribution, where the dose-toxicity parameters of human trials could possibly be drawn, but the new human trial data exclusively to estimate the variance matrix. This means, we split discussions about between-study heterogeneity for the animal datasets and the human datasets. To avoid excessive shrinkage towards the population means for an extreme stratum, we consider a non-exchangeability distribution for each parameter vector that underpins a phase I trial.

5.2 SIGNIFICANCE OF THE WORK

While in statistical literature there exists a number of Bayesian adaptive methods to supplement a new clinical trial with historical data, most rely on the assumption that historical and new trial data are sufficiently similar so that the model parameters can be assumed to be exchangeable. Our research has relaxed this assumption and enabled the possibility of leveraging historical data from heterogeneous data sources. We restrict our attention to the transition step in early drug development to discuss a special type of historical data, which serves as both a motivating and illustrative example for our methodologies. To the best of our knowledge, this thesis comprise three very first proposals to leverage preclinical animal data into human trials given different settings. As phase I clinical trials are generally planned with small sample size, the notorious prior-data conflict can be outstanding than ever.

The Bayesian decision-theoretic approach proposed in Chapter 2 addresses some of the concerns over the sceptisism that whether preclinical data are commensurate enough to be incorporated into a phase I first-in-man trial. This is assessed in a sequential manner, as the clinical trial data accrue. Investigator utilities need to be specified to penalise the inconsistent animal data, which could be identified based on incorrect preclinical predictions of the toxicity in humans. We have provided two forms of a tuning parameter specified to correctly reflect the relevance of the animal data, as our estimator of the overall commensurability between clinical and clinical data can be noisy at early stages of the phase I trial when human toxicity data are sparse. Readers can certainly explore alternatives on their own following this concept. The methodology is suitable to be implemented when animal data are from a single species, while it does not have restriction on the number of animal studies.

The robust Bayesian hierarchical model presented in Chapter 3 suppose animal data are available from a number of preclinical studies performed in different species before the drug is evaluated in humans. Preclinical animal data are translated onto an equivalent human dosing scale, while accounting for uncertainty that surrounds our preliminary knowledge about the intrinsic differences between an animal species and humans. The methodology formalises the process of using historical data that have been recorded in different measure scales. For each animal species, we assume the dose-toxicity parameters of a phase I first-in-man trial to be exchangeable with the parameters of the animal studies, expressed on the common scale. Prior probabilities of exchangeability are formulated to reconcile preclinical information from different animal species. To obtain robust inferences about the dose-toxicity parameters in humans, we have also considered probability of non-exchangeability given a weakly informative prior specifically for the phase I clinical trial. The simulation study was designed to check how our methodology would behave in one of the most extreme scenarios when animal data are accumulated from one species. In an easier scenario when we have animal data from multiple species, better estimates about the between-species heterogeneity can be obtained.

Finally, the generalised Bayesian hierarchical model in Chapter 4 offers a pragmatic solution to integrative subgroup analysis concerned in phase I dose-escalation trials. The issues of data inconsistency could arise between animal species and humans as well as between distinct human subgroups. Our approach promotes efficiency of design and analysis of the new phase I trials while acknowledging heterogeneity in different aspects. They can be undertaken simultaneously or sequentially in different geographic regions. However, none of them will be left on their own devices to draw a conclusion, nor would any trial data override the inferences in another trial. Our Bayesian model works well to achieve the principal goal of estimating region-specific MTDs, while we have noted it cannot be used to describe population-averaged effects of a drug in patients of the same subgroup. This could be possible and we would recommend readers to explore extensions to widen the scope of our investigation.

5.3 limitations and future work

Throughout this thesis, we considered using preclinical animal data, from single or multiple species, to improve efficiency of estimating the MTD(s) for phase I clinical trials, assuming that the historical animal and concurrent human data are generated on the same binary endpoint, i.e., DLT or no DLT, but toxicity data are generally recorded in various dimensions, with different types, grades of severity, attribution and times of occurrence. Phase I dose-escalation trials could be better planned under a new toxicity endpoint paradigm rather than adopting a simplified binary toxicity endpoint. A variaty of toxicity scoring system have been proposed for novel designs for phase I clinical trials; see for example Bekele and Thall [2004]; Yuan et al. [2007]; Cheng et al. [2010]; Ezzalfani et al. [2013]. As one possibility of extending our proposals, investigators may take account of this high dimensional nature of toxicity profile, and more importantly, relate the animal and human toxicity data on each dimension of the toxicity measurement to obtain a more accurate assessment, or preliminary understanding, of the commensurability between preclinical and clinical data that we are interested in.

We have discussed leveraging preclinical animal data exclusively in settings of phase I oncology trials, where efficacy is generally believed to be highly associated with toxicity and therefore the dose-toxicity relationship is of primary interest. In our proposals, we have adopted a two-parameter Bayesian logistic regression model and constrained the activity, specifically, toxicity, of the drug to be monotonically increasing as the dose increases. This works particularly well for cytotoxic agents, whereas we imagine strategies will have to be modified for cancer immunotherapies and molecular targeted agents, as more often than not the dose-efficacy relationship in these fields is unknown and can display a plateau or umbrella shape (Conolly and Lutz [2004]; Jain et al. [2010]). Quite a few trial designs have been proposed to consider both toxicity and efficacy endponits for decision making (Braun [2002]; Nebiyou Bekele and Shen [2005]; Yeung et al. [2015, 2017]; Cai et al. [2014]; Wages and

Tait [2015]; Riviere et al. [2018]; Mozgunov and Jaki [2019]). Extending our Bayesian methodologies to accommodate different endpoints would be interesting, given that collecting preclinical animal data on both toxicity and efficacy endpoints are not uncommon. In accounting for situations concerned with bivariate endpoints, one could evaluate each individually and combine the information through a trade-off function, or jointly model the correlated endpoints based on a copula model (Nelsen [1999]). It would be interesting to explore borrowing of information from animal data, since they could be predictive of either endpoint, or both, of the human trial data to different extent.

In Chapter 4, we have discussed leveraging preclinical data into phase I clinical trials in presence of group heterogeneity. The methodology has been illustrated in the context of phase I bridging trials performed in different geographic regions, for which we do not expect potential ordering between the patient groups. There are situations where we may have quite strong evidence concerning which of the patient group(s) would have higher level of MTD. This could be motivated by the interest of conducting a pediatric phase I trial following adult phase I trials (Smith et al. [1998]; EMA [2017]). We imagine incorporating additional information about the direction and the magnitude of the potential differences between patient groups can improve efficiency of the phase I clinical trials to be designed. EMA [2005b] also suggest to perform non-clinical studies in juvenile animals to predict effects on growth and/or development in the intended age groups, when standard preclinical studies using adult animals and safety data from adult trials cannot well predict toxicity of the drug in all paediatric age groups. We believe this would be worth discussions to make our research more complete.

BIBLIOGRAPHY

Babb, J., Rogatko, A., and Zacks, S. (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine*, 17(10):1103–1120.

Baker, S., Verweij, J., Rowinsky, E., Donehower, R., Schellens, J., Grochow, L., and Sparreboom, A. (2002). Role of body surface area in dosing of investigational anticancer agents in adults, 1991–2001. *JNCI: Journal of the National Cancer Institute*, 94(24):1883–1888.

Balkwill, F., Whitehead, S., Willis, P., Gaymond, N., Kent, A., Page, C., Lovell-Badge, R., Morris, R., Lemon, R., and Banks, D. (2011). Safety of medicines and the use of animals in research. *The Lancet*, 378(9786):127–128.

Balshem, H., MH, Schünemann, H., Oxman, A., Kunz, R., Brozek, J., Vist, G., Falck-Ytter, Y., Meerpohl, J., Norris, S., and Guyatt, G. (2011). Grade guidelines: 3. rating the quality of evidence. *Journal of Clinical Epidemiology*, 64(4):401 – 406.

Bautista, F., Moreno, L., Marshall, L., Pearson, A., Géoerger, B., and Paoletti, X. (2017). Revisiting the definition of dose-limiting toxicities in paediatric oncology phase I clinical trials: An analysis from the innovative therapies for children with cancer consortium. *European Journal of Cancer*, 86:275–284.

Bekele, B. N. and Thall, P. F. (2004). Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *Journal of the American Statistical Association*, 99(465):26–35.

Berry, S. M., Broglio, K. R., Groshen, S., and Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase ii oncology clinical trials. *Clinical Trials*, 10(5):720–734.

Braun, T. M. (2002). The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Controlled Clinical Trials*, 23(3):240 – 256.

Braun, T. M. (2014). The current design of oncology phase I clinical trials: progressing from algorithms to statistical models. *Chinese Clinical Oncology*, 3(1).

Brunier, H. C. and Whitehead, J. (1994). Sample sizes for phase ii clinical trials derived from Bayesian decision theory. *Statistics in Medicine*, 13(23-24):2493–2502.

Cai, C., Yuan, Y., and Ji, Y. (2014). A Bayesian dose finding design for oncology clinical trials of combinational biological agents. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(1):159–173.

Cheng, B., Lee, S. M., and Cheung, Y. K. (2010). Continual reassessment method with multiple toxicity constraints. *Biostatistics*, 12(2):386–398.

Cheung, Y. (2011). *Dose Finding by the Continual Reassessment Method*. Chapman & Hall/CRC Biostatistics Series. Taylor & Francis.

Cheung, Y. K. (2005). Coherence principles in dose-finding studies. *Biometrika*, 92(4):863–873.

Chevret, S. (2006). *Statistical Methods for Dose-Finding Experiments*. Wiley Series in Statistics in Practice. Wiley.

Chow, S. and Liu, J. (2013). *Design and Analysis of Clinical Trials: Concepts and Methodologies*. Wiley Series in Probability and Statistics. Wiley.

Chow, S.-C., Chiang, C., pei Liu, J., and Hsiao, C.-F. (2012). Statistical methods for bridging studies. *Journal of Biopharmaceutical Statistics*, 22(5):903–915.

Conolly, R. B. and Lutz, W. K. (2004). Nonmonotonic Dose-Response Relationships: Mechanistic Basis, Kinetic Modeling, and Implications for Risk Assessment. *Toxicological Sciences*, 77(1):151–157.

Cook, N., Hansen, A., Siu, L., and Abdul Razak, A. (2015). Early phase clinical trials to identify optimal dosing and safety. *Molecular Oncology*, 9(5):997–1007.

Cunanan, K. and Koopmeiners, J. (2017). Hierarchical models for sharing information across populations in phase I dose-escalation studies. *Statistical Methods in Medical Research*, 0(0):1–13.

Curtis, M., Bond, R., Spina, D., Ahluwalia, A., Alexander, S., Giembycz, M., Gilchrist, A., Hoyer, D., Insel, P., Izzo, A., Lawrence, A., MacEwan, D., Moon, L., Wonnacott, S., Weston, A., and McGrath, J. (2015). Experimental design and analysis and their reporting: new guidance for publication in BJP. *British Journal of Pharmacology*, 172(14):3461–3471.

Dane, A. and Wetherington, J. (2014). Statistical considerations associated with a comprehensive regulatory framework to address the unmet need for new antibacterial therapies. *Pharmaceutical Statistics*, 13(4):222–228.

de Haen, P. (1975). The drug lagâdoes it exist in europe? *Drug Intelligence & Clinical Pharmacy*, 9(3):144–150.

Department of Health (2006). *Expert Scientific Group on Phase One Clinical Trials: Final Report*. London: HMSO.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177 – 188.

Diniz, M. A., Quanlin-Li, and Tighiouart, M. (2017). Dose finding for drug combination in early cancer phase I trials using conditional continual reassessment method. *Journal of biometrics biostatistics*, 8(6).

Dresser, R. (2009). First-in-human trial participants: Not a vulnerable population, but vulnerable nonetheless. *The Journal of Law, Medicine & Ethics*, 37(1):38–50.

Duan, Y., Smith, E., and Ye, K. (2006). Using power priors to improve the binomial test of water quality. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(2):151.

Durham, S. D., Flournoy, N., and Rosenberger, W. F. (1997). A random walk rule for phase I clinical trials. *Biometrics*, 53(2):745–760.

Eichler, H., Bloechl-Daum, B., Bauer, P., Bretz, F., Brown, J., Hampson, L., Honig, P., Krams, M., Leufkens, H., Lim, R., Lumpkin, M., Murphy, M., Pignatti, F., Posch, M., Schneeweiss, S., Trusheim, M., and Koenig, F. (2016). "threshold-crossing": a useful way to establish the counterfactual in clinical trials? *Clinical Pharmacology & Therapeutics*, 100(6):699–712.

Eichler, H., Pétavy, F., Pignatti, F., and Rasi, G. (2013). Access to patient-level trial data – a boon to drug developers. *New England Journal of Medicine*, 369(17):1577–1579.

EMA (2005a). *ICH Topic S7B – The Nonclinical Evaluation of the Potential for Delayed Ventricular Repolarization (QT Interval Prolongation) by Human Pharmaceuticals.* European Medicine Agency: London, United Kingdom.

EMA (2005b). *Need for Non-clinical Testing in Juvenile Animals on Human Pharmaceuticals for Paediatric Indications.* European Medicines Agency: London, E14 4HB, UK.

EMA (2008). *Non-clinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorisation for Pharmaceuticals*. European Medicine Agency: London, UK.

EMA (2009). *ICH Topic M3 (R2) – Non-clinical Safety Studies for the Conduct of Human Clinical Trials and Marketing Authorisation for Pharmaceuticals.* European Medicine Agency: London, United Kingdom.

EMA (2011). *ICH Guideline S6 (R1) – Preclinical Safety Evaluation of Biotechnology-driven Pharmaceuticals.* European Medicine Agency: London, United Kingdom.

EMA (2017). *ICH E11 (R1) – Step 5 Guideline on Clinical Investigation of Medical Products in the Pediatric Population.* European Medicines Agency: London, E14 4HB, UK.

Ezzalfani, M., Zohar, S., Qin, R., Mandrekar, S. J., and Deley, M.-C. L. (2013). Dose-finding designs using a novel quasi-continuous endpoint for multipleâtoxicities. *Statistics in Medicine*, 32(16):2728–2746.

FDA (2013). *Guidance for Industry – Preclinical Assessment of Investigational Cellular and Gene Therapy Prducts.* US Food and Drug Administration: Rockville, MD.

Festing, M. and Altman, D. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *Institute for Laboratory Animal Research*, 43(4):244–258.

Fouskakis, D. and Draper, D. (2002). Stochastic optimization: a review. *International Statistical Review*, 70(3):315–349.

French, J., Temkin, N., Shneker, B., Hammer, A., Caldwell, P., and Messenheimer, J. (2012). Lamotrigine xr conversion to monotherapy: first study using a historical control group. *Neurotherapeutics*, 9(1):176–184.

French, J., Wang, S., Warnock, B., and Temkin, N. (2010). Historical control monotherapy design in the treatment of epilepsy. *Epilepsia*, 51(10):1936–1943.

Friede, T., RÃ¶ver, C., Wandel, S., and Neuenschwander, B. (2017). Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biometrical Journal*, 59(4):658–671.

Friedman, L., Furberg, C., and DeMets, D. (1998). *Fundamentals of Clinical Trials*. Springer.

Gamalo-Siebers, M., Savic, J., Basu, C., Zhao, X., Gopalakrishnan, M., Gao, A., Song, G., Baygani, S., Thompson, L., Xia, H. A., Price, K., Tiwari, R., and Carlin, B. P. (2017). Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. *Pharmaceutical Statistics*, 16(4):232–249.

Gasparini, M. and Eisele, J. (2000). A curve-free method for phase I clinical trials. *Biometrics*, 56(2):609–615.

Gasparini, M. and Eisele, J. (2001). Erratum: A curve-free method for phase I clinical trials. *Biometrics*, 57:659–660.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Jakulin, A., Pittau, M., and Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

Gerina-Berzina, A., Vikmanis, U., Teibe, U., and Umbrashko, S. (2012). Anthropometric measurements of the body composition of cancer patients determine the precise role of the body surface area and the calculation of the dose of chemotherapy. *Papers on Anthropology*, 21(0).

Gezmu, M. and Flournoy, N. (2006). Group up-and-down designs for dose-finding. *Journal of Statistical Planning and Inference*, 136(6):1749 – 1764.

GlaxoSmithKline (2016). A study to evaluate the safety, mode of action and clinical efficacy of GSK3050002 in subjects with psoriatic arthritis. *Bethesda (MD): National Library of Medicine (US).* (Available from: https://clinicaltrials.gov/ct2/show/NCT02671188 (NLM Identifier: NCT02671188)) [Last accessed in September 2018].

GlaxoSmithKline (2018). A phase 1, randomized, double-blind (sponsor open), placebo-controlled, single dose escalation trial to evaluate the safety, tolerability pharmacokinetics and pharmacodynamics of GSK3050002 (anti-CCL20 monoclonal antibody) in healthy male volunteers. *GlaxoSmithKline Research & Development Limited*. (Available from: https://www.gsk-clinicalstudyregister.com/files2/gsk-200784-protocol-redact.pdf [Last accessed in September 2018].

Gonnermann, A., Framke, T., Großhennig, A., and Koch, A. (2015). No solution yet for combining two independent studies in the presence of heterogeneity. *Statistics in Medicine*, 34(16):2476–2480.

Gueorguieva, I., Aarons, L., and Rowland, M. (2006). Diazepam pharamacokinetics from preclinical to phase I using a Bayesian population physiologically based pharmacokinetic model with informative prior distributions in winbugs. *Journal of Pharmacokinetics and Pharmacodynamics*, 33(5):571–594.

Guyatt, G., Oxman, A., Kunz, R., Brozek, J., Alonso-Coello, P., Rind, D., Devereaux, P., Montori, V., Freyschuss, B., Vist, G., Jaeschke, R., Williams, J., Murad, M., Sinclair, D., Falck-Ytter, Y., Meerpohl, J., Whittington, C., Thorlund, K., Andrews, J., and Schünemann, H. (2011a). Grade guidelines 6. rating the quality of evidence-imprecision. *Journal of Clinical Epidemiology*, 64(12):1283–1293.

Guyatt, G., Oxman, A., Sultan, S., Glasziou, P., Akl, E., Alonso-Coello, P., Atkins, D., Kunz, R., Brozek, J., Montori, V., Jaeschke, R., Rind, D., Dahm, P., Meerpohl, J., Vist, G., Berliner, E., Norris, S., Falck-Ytter, Y., Murad, M., and Schünemann, H. (2011b). Grade guidelines: 9. rating up the quality of evidence. *Journal of Clinical Epidemiology*, 64(12):1311–1316.

Guyatt, G., Oxman, A., Vist, G., Kunz, R., Brozek, J., Alonso-Coello, P., Montori, V., Akl, E., Djulbegovic, B., Falck-Ytter, Y., Norris, S., Williams, J., Atkins, D., Meerpohl, J., and Schünemann, H. (2011c). Grade guidelines: 4. rating the quality of evidence-study limitations (risk of bias). *Journal of Clinical Epidemiology*, 64(4):407–415.

Hackam, D. and Redelmeier, D. (2006). Translation of research evidence from animals to humans. *JAMA*, 296(14):1727–1732.

Hirst, T., Vesterinen, H., Conlin, S., Egan, K., Antonic, A., Lawson McLean, A., Macleod, M., Grant, R., Brennan, P., Sena, E., and Whittle, I. (2015). A systematic review and meta-analysis of gene therapy in animal models of cerebral glioma: why did promise not translate to human therapy? *Evidence-based Preclinical Medicine*, 1(1):21–33.

Hobbs, B., Carlin, B., Mandrekar, S., and Sargent, D. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056.

Hobbs, B., Carlin, B., and Sargent, D. (2013). Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials*, 10(3):430–440.

Hobbs, B., Sargent, D., and Carlin, B. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*, 7(3):639–674.

Hooijmans, C., de Vries, R., Ritskes-Hoitinga, M., Rovers, M., Leeflang, M., IntHout, J., Wever, K., Hooft, L., de Beer, H., Kuijpers, T., Macleod, M., Sena, E., ter Riet, G., Morgan, R., Thayer, K., Rooney, A., Guyatt, G., SchÃŒnemann, H., Langendam, M., and on behalf of the GRADE Working Group (2018). Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. *PLOS ONE*, 13(1):1–18.

Horton, B. J., Wages, N. A., and Conaway, M. R. (2017). Performance of toxicity probability interval based designs in contrast to the continual reassessment method. *Statistics in Medicine*, 36(2):291–300.

Hsiao, C.-F., Xu, J.-Z., and pei Liu, J. (2004). A two-stage design for bridging studies. *Journal of Biopharmaceutical Statistics*, 15(1):75–83.

Huang, Q., Chen, G., Yuan, Z., and Lan, K. K. G. (2012). Design and sample size considerations for simultaneous global drug development program. *Journal of Biopharmaceutical Statistics*, 22(5):1060–1073.

Iasonos, A., Gounder, M., Spriggs, D. R., Gerecitano, J. F., Hyman, D. M., Zohar, S., and O'Quigley, J. (2012). The impact of non–drug-related toxicities on the estimation of the maximum tolerated dose in phase i trials. *Clinical Cancer Research*, 18(19):5179–5187.

Iasonos, A., Wilton, A. S., Riedel, E. R., Seshan, V. E., and Spriggs, D. R. (2008). A comprehensive comparison of the continual reassessment method to the standard 3 + 3 dose escalation scheme in phase i dose-finding studies. *Clinical Trials*, 5(5):465–477. PMID: 18827039.

Ibrahim, J. and Chen, M. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46–60.

International Conference on Harmonisation (1998). *Tripartite Guidance E5 (R1), Ethnic Factors in the Acceptability of Foreign Clinical Data*. European Medicines Agency: London, E14 4HB, UK.

International Conference on Harmonisation (2006). *Question and Answers for the ICH E5 Guideline on Ethnic Factors in the Acceptability of Foreign Data*. European Medicines Agency: London, E14 4HB, UK.

Jain, R. K., Lee, J. J., Hong, D., Markman, M., Gong, J., Naing, A., Wheler, J., and Kurzrock, R. (2010). Phase I oncology studies: Evidence that in the era of targeted therapies patients on lower doses do not fare worse. *Clinical Cancer Research*, 16(4):1289–1297.

Jaki, T., Clive, S., and Weir, C. (2013). Principles of dose finding studies in cancer: a comparison of trial designs. *Cancer Chemotherapy and Pharmacology*, 71(5):1107–1114.

Ji, Y., Liu, P., Li, Y., and Bekele, B. N. (2010). A modified toxicity probability interval method for dose-finding trials. *Clinical Trials*, 7(6):653–663.

John, W., Helene, T., and Anne, W. (2010). A bayesian dose-finding procedure for phase I clinical trials based only on the assumption of monotonicity. *Statistics in Medicine*, 29(17):1808–1824.

José, P. and Stephen, D. (2009). Exposure response â getting the dose right. *Pharmaceutical Statistics*, 8(3):173–175.

Knapp, G. and Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22(17):2693–2710.

Kouno, T., Katsumata, N., Mukai, H., Ando, M., and Watanabe, T. (2003). Standardization of the body surface area (bsa) formula to calculate the dose of anticancer agents in Japan. *Japanese Journal of Clinical Oncology*, 33(6):309–313.

Le Tourneau, C., Lee, J. J., and Siu, L. L. (2009). Dose escalation methods in phase I cancer clinical trials. *Journal of the National Cancer Institute*, 101(10):708–720.

Le Tourneau, C., Razak, A., Gan, H., Pop, S., Dieras, V., Tresca, P., and Paoletti, X. (2011). Heterogeneity in the definition of dose-limiting toxicity in phase I cancer clinical trials of molecularly targeted agents: A review of the literature. *European Journal of Cancer*, 47(10):1468–1475.

Leon-Novelo, L., Bekele, B., Müller, P., Quintana, F., and Wathen, K. (2012). Borrowing strength with nonexchangeable priors over subpopulations. *Biometrics*, 68(2):550–558.

Li, N. and Wang, W. (2012). Practical and statistical considerations on simultaneous global drug development. *Journal of Biopharmaceutical Statistics*, 22(5):1074–1077.

Lin, Y. and Shih, W. J. (2001). Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. *Biostatistics*, 2(2):203–215.

Lindley, D. (1972). *Bayesian Statistics, a review.* PA: SIAM. (A sharp comprehensive reivew of the whole subject up to the 1970sâ, emphasising its internal consistency).

Liu, S., Pan, H., Xia, J., Huang, Q., and Yuan, Y. (2015). Bridging continual reassessment method for phase i clinical trials in different ethnic populations. *Statistics in Medicine*, 34(10):1681–1694.

Liu, S. and Yuan, Y. (2015). Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(3):507–523.

Love, S. B., Brown, S., Weir, C. J., Harbron, C., Yap, C., Gaschler-Markefski, B., Matcham, J., Caffrey, L., McKevitt, C., Clive, S., Craddock, C., Spicer, J., and Cornelius, V. (2017). Embracing model-based designs for dose-finding trials. *British journal of cancer*, 117(3):332â339.

Lunn, D., Barrett, J., Sweeting, M., and Thompson, S. (2013). Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(4):551–572.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The bugs project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067.

Maccario, J., O'quigley, J., and Paoletti, X. (2002). Nonâparametric optimal design in dose finding studies. *Biostatistics*, 3(1):51–56.

Macleod, M. R., Lawson McLean, A., Kyriakopoulou, A., Serghiou, S., de Wilde, A., Sherratt, N., Hirst, T., Hemblade, R., Bahor, Z., Nunes-Fonseca, C., Potluru, A., Thomson, A., Baginskitae, J., Egan, K., Vesterinen, H., Currie, G. L., Churilov, L., Howells, D. W., and Sena, E. S. (2015). Risk of bias in reports of in vivo research: A focus for improvement. *PLOS Biology*, 13(10):1–12.

Mandrekar, S. and Sargent, D. (2009). Clinical trial designs for predictive biomarker validation: One size does not fit all. *Journal of Biopharmaceutical Statistics*, 19(3):530–542. PMID: 19384694.

Mizugaki, H., Yamamoto, N., Fujiwara, Y., Nokihara, H., Yamada, Y., and Tamura, T. (2015). Current status of single-agent phase i trials in japan: Toward globalization. *Journal of Clinical Oncology*, 33(18):2051–2061.

Morita, S. (2011). Application of the continual reassessment method to a phase i dose-finding trial in japanese patients: East meets west. *Statistics in Medicine*, 30(17):2090–2097.

Morita, S., Thall, P. F., and Müller, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602.

Mozgunov, P. and Jaki, T. (2019). An information theoretic phase iâii design for molecularly targeted agents that does not require an assumption of monotonicity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(2):347–367.

Muller, P., Berry, D. A., Grieve, A. P., and Krams, M. (2006). A Bayesian decision-theoretic dose-finding trial. *Decision Analysis*, 3(4):197–207.

Müller, P. and Mitra, R. (2013). Bayesian nonparametric inference – why and how. *Bayesian Analysis*, 8(2):269–302.

National Cancer Institute (2017). *Common terminology criteria for adverse events Version 5.0*. US Department of Health and Human Services.

Nebiyou Bekele, B. and Shen, Y. (2005). A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. *Biometrics*, 61(2):343–354.

Nelsen, R. (1999). *An Introduction to Copulas (Springer Series in Statiatics)*. New York: Springer.

Neuenschwander, B., Branson, M., and Gsponer, T. (2008). Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in Medicine*, 27(13):2420–2439.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566.

Neuenschwander, B., Capkun-Niggli, G., Branson, M., and Spiegelhalter, D. (2010). Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18.

Neuenschwander, B., Roychoudhury, S., and Schmidli, H. (2016a). On the use of co-data in clinical trials. *Statistics in Biopharmaceutical Research*, 8(3):345–354.

Neuenschwander, B., Wandel, S., Roychoudhury, S., and Bailey, S. (2016b). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics*, 15(2):123–134.

Ogura, T., Morita, S., Yonemori, K., Nonaka, T., and Urano, T. (2014). Exploring ethnic differences in toxicity in early-phase clinical trials for oncology drugs. *Therapeutic Innovation & Regulatory Science*, 48(5):644–650.

O'Quigley, J. (2002). Curve-free and model-based continual reassessment method designs. *Biometrics*, 58(1):245–249.

O'Quigley, J. and Iasonos, A. (2014). Bridging solutions in dose finding problems. *Statistics in biopharmaceutical research*, 6(2):185â197.

O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*, 46(1):33–48.

O'Quigley, J. and Shen, L. Z. (1996). Continual reassessment method: A likelihood approach. *Biometrics*, 52(2):673–684.

O'Quigley, J., Shen, L. Z., and Gamst, A. (1999). Two-sample continual reassessment method. *Journal of Biopharmaceutical Statistics*, 9(1):17–44.

Paoletti, X. and Doussau, A. (2014). Dose finding methods in oncology: From the maximum tolerated dose to the recommended phase II dose. In van Montfort, K., Oud, J., and Ghidey, W., editors, *Developments in Statistical Evaluation of Clinical Trials*, chapter 18. Springer Berlin Heidelberg.

Paoletti, X. and Kramar, A. (2009). A comparison of model choices for the continual reassessment method in phase I cancer trials. *Statistics in Medicine*, 28(24):3012–3028.

Pharmaceuticals and Medical Devices Agency in Japan (2007). *Basic Principles on Global Clinical Trials, Notification No. 0928010*. Ministry of Health, Labour and Welfare: Tokyo, Japan. This document is an informal translation by PMDA of the final notification published in Japanese on Sep. 28th 2007 and is intended to use as a reference for considering global clinical trials. Last access: September, 2018.

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29(3):175 – 188.

R Core Team (2017). *R: Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reigner, B. and Blesch, K. (2001). Estimating the starting dose for entry into humans: principles and practice. *European Journal of Clinical Pharmacology*, 57:835–845.

Rhodes, K., Turner, R., White, I., Jackson, D., Spiegelhalter, D., and Higgins, J. (2016). Implementing informative priors for heterogeneity in meta-analysis using meta-regression and pseudo data. *Statistics in Medicine*, 35(29):5495–5511.

Riviere, M.-K., Yuan, Y., Jourdan, J.-H., Dubois, F., and Zohar, S. (2018). Phase I/II dose-finding design for molecularly targeted agent: Plateau determination using adaptive randomization. *Statistical Methods in Medical Research*, 27(2):466–479.

Roberts, I., Kwan, I., Evans, P., and Haig, S. (2002). Does animal experimentation inform human healthcare? observations from a systematic review of international animal experiments on fluid resuscitation. *British Medical Journal*, 324(7335):474–476.

Rogge, M. and Taft, D. (2016). *Preclinical Drug Development*. Drugs and the Pharmaceutical Sciences. CRC Press.

Roman, D., VerHoeve, J., Schadt, H., Vicart, A., Walker, U., Turner, O., Richardson, T., Wolford, S., Miller, P., Zhou, W., Lu, H., Akimov, M., and Kluwe, W. (2016). Ocular toxicity of auy922 in pigmented and albino rats. *Toxicology and Applied Pharmacology*, 309(Supplement C):55–62.

Röver, C., Knapp, G., and Friede, T. (2015). Hartung-knapp-sidik-jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology*, 15(1):99.

Saville, B., Connor, J., Ayers, G., and Alvarez, J. (2014). The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clinical Trials*, 11(4):485–493. PMID: 24872363.

Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D., and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032.

Sessa, C., Shapiro, G. I., Bhalla, K. N., Britten, C., Jacks, K. S., Mita, M., Papadimitrakopoulou, V., Pluard, T., Samuel, T. A., Akimov, M., Quadt, C., Fernandez-Ibarra, C., Lu, H., Bailey, S., Chica, S., and Banerji, U. (2013). First-in-human phase i dose-escalation study of the hsp90 inhibitor auy922 in patients with advanced solid tumors. *Clinical Cancer Research*, 19(13):3671–3680.

Sharma, V. and McNeill, J. (2009). To scale or not to scale: the principles of dose extrapolation. *British Journal of Pharmacology*, 157(6):907–921.

Simon, R. J., Freidlin, B., Rubinstein, L., Arbuck, S. G., Collins, J. L., and Christian, M. (1997). Accelerated titration designs for phase I clinical trials in oncology. *Journal of the National Cancer Institute*, 89(15):1138–1147.

Smith, M., Bernstein, M., Bleyer, W. A., Borsi, J. D., Ho, P., Lewis, I. J., Pearson, A., Pein, F., Pratt, C., Reaman, G., Riccardi, R., Seibel, N., Trueworthy, R., Ungerleider, R., Vassal, G., and Vietti, T. (1998). Conduct of phase I trials in children with cancer. *Journal of Clinical Oncology*, 16(3):966–978.

Stallard, N. and Todd, S. (2011). Seamless phase ii/iii designs. *Statistical Methods in Medical Research*, 20(6):623–634.

Stephen, S., Dipti, A., A., B. R., M., B. S., Barbara, B., Peter, C., Andrew, G., Andrew, G., and Peter, L. (2007). Statistical issues in first-in-man studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(3):517–579.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics*, 45(3):925–937.

Sutton, A. and Abrams, K. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10(4):277–303.

Takeda, K. and Morita, S. (2018). Bayesian dose-finding phase I trial design incorporating historical data from a preceding trial. *Pharmaceutical Statistics*, 0(0):1–11.

Tang, L., Persky, A., Hochhaus, G., and Meibohm, B. (2004). Pharmacokinetic aspects of biotechnology products. *Journal of Pharmaceutical Sciences*, 93(9):2184–2204.

Thall, P. F. and Wathen, J. K. (2007). Practical bayesian adaptive randomisation in clinical trials. *European Journal of Cancer*, 43(5):859 – 866.

Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H., and Benjamin, R. S. (2003). Hierarchical bayesian approaches to phase II trials in diseases with multiple subtypes. *Statistics in Medicine*, 22(5):763–780.

Thomas, N. (2017). *R2OpenBUGS: Running OpenBUGS from R*. R package version 3.2.

Tsong, Y. (2012). Statistical considerations on design and analysis of bridging and multiregional clinical trials. *Journal of Biopharmaceutical Statistics*, 22(5):1078–1080.

Tsou, H.-H., Tsong, Y., Liu, J.-T., Dong, X., and Wu, Y. (2012). Weighted evidence approach of bridging study. *Journal of Biopharmaceutical Statistics*, 22(5):952–965.

Turner, R., Jackson, D., Wei, Y., Thompson, S., and Higgins, J. (2015). Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine*, 34(6):984–998.

Ueno, T., Asahina, Y., Tanaka, A., Yamada, H., Nakamura, M., and Uyama, Y. (2013). Significant differences in drug lag in clinical development among various strategies used for regulatory submissions in japan. *Clinical Pharmacology & Therapeutics*, 95(5):533–541.

USFDA (2003). *Exposure-Response Relationships – Study Design, Data Analysis, and Regulatory Applications*. US Food and Drug Administration: Rockville, MD.

USFDA (2005). *Estimating the maximum safe starting dose in initial clinical trials for therapeutics in adult healthy volunteers*. US Food and Drug Administration: Rockville, MD.

van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B., and Lesaffre, E. (2017). Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical Methods in Medical Research*, 0(0):1–16.

Vehtari, A. and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statist. Surv.*, 6:142–228.

Vehtari, A. and Ojanen, J. (2014). Errata: A survey of bayesian predictive methods for model assessment, selection and comparison. *Statist. Surv.*, 8:1.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., and Thompson, L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54.

Viergever, R. F. and Li, K. (2015). Trends in global clinical trial registration: an analysis of numbers of registered clinical trials in different parts of the world from 2004 to 2013. *BMJ Open*, 5(9).

Wadsworth, I., Hampson, L., and Jaki, T. (2018). Extrapolation of efficacy and other data to support the development of new medicines for children: a systematic review of methods. *Statistical Methods in Medical Research*, 27(2):398–413.

Wages, N. A., Read, P. W., and Petroni, G. R. (2015). A phase i/ii adaptive design for heterogeneous groups with application to a stereotactic body radiation therapy trial. *Pharmaceutical Statistics*, 14(4):302–310.

Wages, N. A. and Tait, C. (2015). Seamless phase I/II adaptive design for oncology trials of molecularly targeted agents. *Journal of Biopharmaceutical Statistics*, 25(5):903–920.

Weber, J. S., Levit, L. A., Adamson, P. C., Bruinooge, S., Burris, H. A., Carducci, M. A., Dicker, A. P., GÃ¶nen, M., Keefe, S. M., Postow, M. A., Thompson, M. A., Waterhouse, D. M., Weiner, S. L., and Schuchter, L. M. (2015). American society of clinical oncology policy statement update: The critical role of phase i trials in cancer research and treatment. *Journal of Clinical Oncology*, 33(3):278–284.

West, G. and Brown, J. (2005). The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. *Journal of Experimental Biology*, 208(9):1575–1592.

Whitehead, J. (2006). Using bayesian decision theory in dose-escalation studies. In Chevret, S., editor, *Statistical Methods for Dose-Finding Experiments*, Statistics in Practice, chapter 7. Wiley-Blackwell.

Whitehead, J. and Williamson, D. (1998). Bayesian decision procedures based on logistic regression models for dose-finding studies. *Journal of Biopharmaceutical Statistics*, 8(3):445–467.

Wileman, H. and Mishra, A. (2010). Drug lag and key regulatory barriers in the emerging markets. *Perspectives in Clinical Research*, 1(2):51–56.

World Medical Association (2013). World medical association declaration of Helsinki: Ethical principles for medical research involving human subjects. *Journal of American Medical Association*, 310(20):2191–2194.

Yeung, W. Y., Reigner, B., Beyer, U., Diack, C., SabanÃ©s bovÃ©, D., Palermo, G., and Jaki, T. (2017). Bayesian adaptive dose-escalation designs for simultaneously estimating the optimal and maximum safe dose based on safety and efficacy. *Pharmaceutical Statistics*, 16(6):396–413.

Yeung, W. Y., Whitehead, J., Reigner, B., Beyer, U., Diack, C., and Jaki, T. (2015). Bayesian adaptive dose-escalation procedures for binary and continuous responses utilizing a gain function. *Pharmaceutical Statistics*, 14(6):479–487.

Yin, G. and Yuan, Y. (2009). Bayesian model averaging continual reassessment method in phase i clinical trials. *Journal of the American Statistical Association*, 104(487):954–968.

Yuan, Z., Chappell, R., and Bailey, H. (2007). The continual reassessment method for multiple toxicity grades: A Bayesian quasi-likelihood approach. *Biometrics*, 63(1):173–179.

Zhang, T., Lipkovich, I., and Marchenko, O. (2017). Bridging data across studies using frequentist and bayesian estimation. *Journal of Biopharmaceutical Statistics*, 27(3):426–441.

Zhou, Y. and Whitehead, J. (2003). Practical implementation of Bayesian dose-escalation procedures. *Drug Information Journal*, 37(1):45–59.

Zohar, S. and Chevret, S. (2007). Recent developments in adaptive designs for phase I/II dose-finding studies. *Journal of Biopharmaceutical Statistics*, 17(6):1071–1083.