Bayesian methods for borrowing information in clinical drug development

Dissertation

zur Erlangung des humanwissenschaftlichen Doktorgrades

in der Medizin

der Georg-August-Universität Göttingen

vorgelegt von

Burak Kürşad GÜNHAN

aus der Türkei

Göttingen, 2020

Supervisor:	Prof. Dr. Tim Friede Institut für Medizinische Statistik Universitätsmedizin Göttingen
Second Committee Member:	Prof. Dr. Thomas Kneib Professur für Statistik und Ökonometrie Georg-August-Universität Göttingen
Third Committee Member:	Prof. Dr. Markus Zabel Klinik für Kardiologie und Pneumologie Universitätsmedizin Göttingen
Day of Disputation:	07.12.2020

Declaration of Authorship

I, Burak Kürşad GÜNHAN, declare that this dissertation titled, "Bayesian methods for borrowing information in clinical drug development" and the work presented in it are my own and that it was written independently with no other sources and aids than quoted.

"A Bayesian is one who asks you what you think before a clinical trial in order to tell you what you think afterwards."

Stephen Senn

"It is not innovative, if it does not work."

Anonymous

Abstract

Clinical drug development is the process of investigating potential pharmaceutical therapies in clinical trials. The clinical investigation of drugs consists of four phases. The aim is bringing a candidate drug from early phase trials to a product approved for public use by the drug regulatory agencies. Clinical drug development is highly expensive and highly time consuming enterprise with very low probability of success. Thus, increasing the efficiency of clinical trials is critical, especially in the early phases of clinical drug development.

One way to improve efficiency of the clinical trials is to utilize (or borrow) relevant information from external sources. Bayesian statistics is the mathematical procedure to update our prior distributions of the unknown parameters given the available data. The recursive nature of Bayesian statistics provides a promising framework for borrowing information. Another advantage of Bayesian statistics is to enable us to built more complicated models with the help of Markov chain Monte Carlo computation techniques. This helps to include, for instance, hierarchical structures in the model, when they are supported by the data. However, complicated models must be calibrated well, especially in the presence of sparse data, such as in early phase trials.

The first aim of this dissertation is to investigate phase I trials involving multiple treatment schedules. A treatment schedule refers to a frequency of administration. There are two possible types of such trials: simultaneous and sequential investigations of multiple schedules. In a simultaneous design, doses and schedules are varied simultaneously in the same trial. In a sequential design, the information from a completed phase I design stage of a trial is used to inform a new phase I design stage with a different treatment schedule. To design and analyze both types of trials, I develop a Bayesian time-to-event pharmacokinetic (TITE-PK) model. The developed model uses PK principles to borrow information from different treatment schedules explicitly. Furthermore, TITE-PK makes use of an adapted escalation-with-overdose-control criterion to control the number of patients administered with overly toxic doses. For both types of investigations of multiple schedules, simulation results of TITE-PK yield desirable performance in terms of the common metrics such as the correct maximum tolerated dose declarations and the mean number of required patients in the trial.

The second aim of this dissertation is to investigate phase II dose-finding trials involving multiple schedules, which is motivated by a phase II trial in atopic dermatitis. A common approach to estimate the dose-response function in such trials is pooling doses from different schedules after re-scaling them based on the frequency of administration. Recently, a partial pooling approach has been suggested, in which certain parameters are treated as schedule specific fixed-effects. As an alternative, I

propose to use a Bayesian hierarchical model in which certain parameters are treated as random-effects, while others are assumed to be shared between schedules. Estimates of the dose-response function for each schedule are obtained by borrowing. In simulations, the proposed method yields better performance compared to complete pooling and partial pooling with fixed-effects in terms of the investigated metrics such as the mean absolute error and the mean coverage probability of interval estimates. I develop a publicly available R package, `ModStan`, to automate the implementation.

The third aim of this dissertation is to study meta-analyses of few studies involving rare safety events. Meta-analysis is using statistical methods to combine multiple trials. Trials with no or very rare events, which can produce considerable bias in the estimation, are a major challenge. To overcome this, I suggest the use of a weakly informative prior (WIP) for the treatment effect parameter in a binomial-normal hierarchical model as a penalization technique. A WIP is constructed by assuming a normal prior with zero mean and an a priori interval for plausible values. Furthermore, the suggested WIP is verified empirically using the Cochrane Database of Systematic Reviews. The proposed method is assessed in simulations. It displays better or similar performance in terms of the accuracy of point estimates and the coverage probability of interval estimates compared to standard methods. The proposed method is illustrated by a meta-analysis dataset in pediatric transplantation. I implement the proposed method as a publicly available R package, `MetaStan`.

Acknowledgements

I take this opportunity to express my gratitude and appreciation to all people who make this dissertation possible.

I had the great opportunity of working under the supervision of Prof. Tim Friede, who encouraged me to pursue research in this PhD topic. I would like to thank him for his continuous support and motivation. I am very grateful to him for his suggestions and guidance of my research, for very fruitful discussions, and for providing funding of my work.

Also, I would like to thank my thesis committee members, Prof. Thomas Kneib and Prof. Markus Zabel, for their input. I am very grateful to Dr. Christian Röver for his willingness to discuss my work, his open-door policy and collaborating on one of my papers.

I would like to express my gratitude to Dr. Sebastian Weber and Abdelkader Seroutou for being co-authors in my papers and mentoring me during my internship at Novartis Pharma AG in Basel. Furthermore, I would like to thank Paul Meyvisch for collaborating on one of my papers and the supervision during my internship at Galapagos NV in Mechelen.

In addition to my advisors, I won't forget to express my gratitude to my fellow PhD students and the staff at our department for simulating discussions and the fun-time we spent together. In particular, I am grateful to Cynthia for sharing a very nice working environment on the top floor, and Christian, Roland and Markus for the nice after-lunch discussion.

I would like to thank Tobias, Christian and Cynthia for proofreading this thesis.

In addition, I would like to thank my parents, my sister, and my friends, in particular Alaa and Faizal, for their immense support, encouragement and motivation.

Contents

To my sister, Tuba

1 Introduction

1.1 Phases of clinical drug development

Drug development is the process of investigating pharmaceutical therapies which may eventually be used as therapies. After the preclinical phase is completed, the clinical investigation of drugs consists of four phases: namely phases I, II, III, and IV (Chuang-Stein and Kirby, 2017, Chapter 1). Phase I or first-in-human trials are conducted to assess the PK profile and the safety of the potential therapy. In phase II trials, the relationship between the dose and the response of the drug are investigated. Phase III trials are confirmatory trials, which aim to determine the efficacy and safety of the drug. Typically, larger numbers of patients are recruited for phase III trials compared to the previous phases. After regulatory approval, phase IV trials are performed in order to evaluate rare and/or long-term risks. Since I focus on phase I and II trials in this dissertation, I introduce them in more detail in the following.

In phase I trials, the main purpose of phase I trials is to determine the maximum tolerated dose (MTD), that is, the highest dose without generating unacceptable adverse effects. The trial is, usually, conducted through dose-escalation steps, hence it is called a phase I dose-escalation trial. The observed adverse effects are classified into dose-limiting toxicities (DLT) and non-DLT. Based on the number of DLT from small cohorts of patients, the dose recommendation for the next cohort is determined (Le Tourneau, Lee, and Siu, 2009). Because the identification of the MTD will influence the final product, a reliable and efficient MTD identification is essential for drug development (Le Tourneau, Lee, and Siu, 2009). In phase II dose-finding trials, patients usually are randomized to different dose levels or the control, or sometimes more than one control, e.g. placebo and active control. There are two main goals: (1) establishing the dose-response signal and (2) characterizing the dose-response relationship (Ruberg, 1995). Since the understanding of the dose-response relationship is fundamental to the drug development, phase II trials are a crucial part of development program.

Any treatment plan of a phase I or II trial includes the amount of dose given to the patients and how often it will be given, in other words, a treatment schedule. A treatment schedule can be defined as the frequency of administration within a treatment cycle, for instance a weekly or a daily schedule. Traditionally, statistical methods for phase I or II trials involve merely one treatment schedule. In recent years, multiple schedules are investigated in phase I or II trials, for instance in oncology (de Lima et al., 2010; Besse et al., 2014), in atopic dermatitis (Thaçi et al., 2016; Galapagos NV, 2020), or in hypercholesterolemia (Pfizer, 2017). In such trials, the identification of MTD in a phase I trial or the dose-response relationship in a phase II trial may also be influenced by the schedule. However, very limited literature

exists on the the statistical methods for phase I and II trials involving multiple treatment schedules (Guo, Li, and Yuan, 2016). For instance, it has been shown in some clinical trials that the toxicity (Shah et al., 2008) or the efficacy (Wagner et al., 2010) may depend on the treatment schedule. These suggest that the safety and/or efficacy of a drug is not only a function of the dose, but also the schedule. Therefore, different types of statistical methods are required to make use of relevant information, in other words to reliably borrow information from different schedules (Wages, 2017).

1.2 Meta-analysis

Meta-analysis is using statistical methods to combine multiple studies to address a question of interest, for example to estimate the safety profile of a drug. Popular meta-analysis models include a fixed-effect model and a random-effects model, for instance DerSimonian-Laird method (DerSimonian and Laird, 1986). In the latter, the potential heterogeneity in treatment effects between trials are taken into account. The heterogeneity may stem from the fact that trials can have some important differences, e.g., different patient populations or types of administrations of drugs etc. Hence, random-effects models are regarded as more reliable for many application areas (Higgins, Thompson, and Spiegelhalter, 2009).

When a safety event is not specified as the primary outcome in a trial and include sample size accordingly, the trial may not have enough subjects to detect safety events (U.S. Food and Drug Administration, 2018). Thus, for such trials, a meta-analysis of clinical trials is of greater importance (Higgins and Green, 2008, Chapter 10). The Cochrane Database of Systematic Reviews (CDSR) is one of the most comprehensive databases of systematic reviews in health care (Cochrane, 2020). Based on a random sample of meta-analyses from the CDSR, more than 50% of meta-analyses of safety events contained trials with event probabilities smaller than 5% (Vandermeer et al., 2009). Thus, safety events are typically rare events meaning that zero or a very small number of events are observed in the case of count outcomes. Since popular meta-analysis methods, for instance DerSimonial-Laird method, are suitable for the meta-analysis of common events, special methods for the meta-analysis of rare events are required. When the number of studies which are meta-analyzed is small (five or less), the investigation of the potential heterogeneity between trials becomes harder (Gelman, 2006). Based on an analysis of the CDSR, half of the meta-analyses from the CDSR include two or three trials (Turner, Davey, et al., 2012). Thus, statistical models for random-effects meta-analysis of few studies involving rare events is required to reliably obtain an overall treatment effect for the outcome of interest.

1.3 Bayesian methods in clinical drug development

In Bayesian statistics, model parameters are treated as random quantities, while data is treated as fixed. This is in contrast to the frequentist statistics, in which treatment is the other way around. The basis for Bayesian statistics is the Bayes' theorem. The

posterior distribution of the unknown parameter θ given the observed data x is

$$f(\theta|x) \approx f(x|\theta)\, f(\theta)$$

where $f(x|\theta)$ is the likelihood function and $f(\theta)$ is the prior distribution (Gelman, Carlin, et al., 2013, Chapter 1). The necessity to specify a prior distribution for model parameters is another important difference between Bayesian and frequentist statistics. Statistical inference in Bayesian statistics is based on the posterior distribution. However, computing the posterior distributions of the parameters can be a challenge. Since analytical approaches are limited to some set of prior distributions and likelihood functions, stochastic approximations such as Markov chain Monte Carlo (MCMC) methods (Gelman, Carlin, et al., 2013, Chapter 11) are commonly used. In the dissertation, I use a modern MCMC engine called **Stan**, which employs the Hamiltonian Monte Carlo with No-U-Turn Sampler (Carpenter et al., 2017).

Bayesian statistical methods are commonly used in the early phases of drug development (Price and LaVange, 2014). One of the first uses of Bayesian methods goes back to the Continual Reassessment Method (CRM) for phase I dose-escalation trials by O'Quigley, Pepe, and Fisher (1990). In the past years, many early phase trials used Bayesian methods as the primary analysis, including phase I trials (Demetri et al., 2009; de Lima et al., 2010; Angevin et al., 2013; Besse et al., 2014; Esaki et al., 2019; Naing et al., 2020) and phase II trials (Baeten et al., 2013; Murphy et al., 2017; Amgen, 2019). In addition to the early phase trials, Bayesian methods are recommended by the regulatory agencies for trials involving rare diseases (European Medicines Agency, 2006; U.S. Food and Drug Administration, 2014). Additionally, U.S. Food and Drug Administration issued a guideline for the use of Bayesian statistics in medical device clinical trials (U.S. Food and Drug Administration, 2010). Bayesian methods are also mentioned in the guidelines for adaptive designs for clinical trials (U.S. Food and Drug Administration, 2019). For the meta-analysis of clinical trials, Bayesian approaches have been suggested, especially for the random-effects meta-analysis of few trials (Smith, Spiegelhalter, and A. Thomas, 1995; Sutton and Abrams, 2001). In the guideline for meta-analysis of clinical trials issued by the U.S. Food and Drug Administration, Bayesian methods are mentioned as an alternative to frequentist methods (U.S. Food and Drug Administration, 2018).

Bayesian statistics provides a suitable framework to exploit similarity and borrow relevant information across strata, for example across schedules in a phase II dose-finding trial (Viele et al., 2014). Borrowing information can be implemented in meta-analysis models by investigating the corresponding shrinkage estimate (Röver and Friede, 2020). Borrowing information can improve the frequentist properties, including the accuracy of the point estimates and the coverage probabilities of the interval estimates.

1.4 Research questions

In this dissertation, I propose Bayesian methods for clinical drug development and for meta-analyses in the sparse data situations. The application areas include phase I dose-escalation trials and phase II dose-finding trials, in which the trials are of small to moderate size, thus borrowing information is crucial. Furthermore, I consider

meta-analysis of rare events, for instance safety events, where the individual studies usually do not include enough subjects to detect differences between event rates. These three areas are introduced in three sections.

1.4.1 Phase I dose-escalation trials with multiple schedules

There are two different designs of phase I dose-escalation trials with multiple schedules, namely simultaneous and sequential designs. In the simultaneous design, the dose and schedule are varied simultaneously, and a dose-schedule combination is recommended for the next cohort of patients. Finally, the maximum tolerated dose and schedule combination (MTC) is calculated. The sequential design consists of k design stages, which equals the number of schedules investigated in the trial. Assume that the schedules are denoted by S_i where $i = 1, 2, \ldots, k$. Patients are administered with the schedule S_1, and the maximum tolerated dose (MTD) is declared for S_1 in the first step. Then, patients are administered with schedule S_2, the decision for the next cohort is based on the data from both schedules S_1 (completed trial) and S_2 (ongoing trial). Finally, the MTD is declared for the schedule S_2. This sequential approach can continue for schedule S_3 and so on.

The literature on simultaneous designs is very limited (Wages, 2017). One example of a simultaneous investigation of multiple schedules in a phase I dose-escalation trial is the Vidaza trial (ClinicalTrials.gov identifier: NCT01080664) (de Lima et al., 2010). Vidaza is a cytotoxic drug used to treat asmyelodysplastic syndrome, a blood cell disease. In the Vidaza trial, four different schedules and three doses are investigated simultaneously. A Bayesian time-to-event model (Braun et al., 2007) was used to determine dose and schedule decisions in the trial. However, this model requires approximately 60 patients as the total sample size, which is not feasible for many trials. Alternatively, Wages, O'Quigley, and Conaway (2014) introduced the partial ordered continual reassessment method (POCRM), which has been shown to require approximately 25 patients. Furthermore, POCRM relaxes the assumption of complete ordered schedules, that is, DLT probabilities increase with schedules of more frequent administration given the same cumulative dose. Other methods developed for the simultaneous design jointly model efficacy and toxicity (Thall et al., 2013; Guo, Li, and Yuan, 2016; Cunanan and Koopmeiners, 2017), thus they have a different focus.

A sequential phase I trial with multiple strata, where strata may refer to subpopulations, route of administrations, or treatment schedules is called a bridging phase I trial. Different statistical methods to analyze bridging trials include the bridging continual reassessment method (B-CRM) (Liu et al., 2015), the Bayesian Logistic Regression Model using a meta-analytic-predictive prior (Neuenschwander, Matano, et al., 2015), and the continual reassessment method using an adaptive power prior approach (Ollier et al., 2020).

As opposed to the existing methods, we explicitly model different treatment schedules by considering an exposure-response model instead of a dose-response model. The exposure measure of the drug is calculated using a pharmacokinetic (PK) model in which the frequency of administration is taken into account. To this end, I developed a Bayesian time-to-event pharmacokinetics (TITE-PK) model. TITE-PK models time-to-first DLTs using the planned schedule in a fully Bayesian framework.

TITE-PK can be used for dose-schedule decisions of both simultaneous and sequential designs of phase I dose-escalation trials with multiple schedules.

1.4.2 Phase II dose-finding trials with multiple schedules

A simple approach to analyze a phase II trial with multiple schedule is estimating separate dose-response functions for each schedule. However, this method ignores potential similarities of dose-response functions between schedules. As an alternative, one can re-scale all doses by converting the doses into a reference schedule using their corresponding frequency of administrations. Hence, one can pool all re-scaled doses to estimate the dose-response functions of different schedules. This is considered as complete pooling. For instance, if the trial involve weekly and monthly schedules, the complete pooling assumes that x mg dose with a monthly schedule equals to $x/2$ mg dose with a weekly schedule. Thus, this method does not take into account potential heterogeneity in dose-response models between schedules.

A middle ground between two approaches above is assuming some parameters of the dose-response model are shared between schedules, while others are allowed to be different, that is, partial pooling proposed by Feller et al. (2017) and Möllenhoff, Bretz, and Dette (2020). They proposed to treat the unshared parameters as schedule specific fixed-effects in a partial pooling approach. To borrow information between schedules, I propose to use schedule specific random-effects for some parameters of the dose-response function, while others are assumed to be shared. Shrinkage estimation is used to obtain schedule specific random-effects for certain parameters. It has been shown that shrinkage estimates improve the long-run properties of the estimates, for instance the mean squared error, in comparison to a stratified analysis or complete pooling (Neuenschwander, Wandel, et al., 2016). Shrinkage estimation has been proposed for clinical trials with multiple strata, for instance to estimate a response rate in the presence of multiple patient populations (Jones et al., 2011). Here, I consider shrinkage estimates of certain parameters of a dose-response model. To the best of my knowledge, this has not been investigated. The proposed method is illustrated using a phase II trial with multiple schedules in atopic dermatitis. The long-run properties including the mean bias of the dose-response function are studied in simulations.

1.4.3 Meta-analysis of rare events with few studies

Data sparsity in meta-analysis is commonly reflected by the number of studies with no events either in one arm (single-zero study) or in both arms (double-zero study). Furthermore, the data sparsity problem is amplified, when the number of studies included in a meta-analysis is low. Standard meta-analysis methods rely on large-sample properties. For example, they usually depend on the computation of individual log-odds ratio estimates, which are not available in case of single-zero or double-zero studies. Therefore, they are not very suitable to conduct meta-analyses of few studies involving rare events (Bradburn et al., 2007). Many methods have been proposed for the meta-analysis of rare events in the literature, including the

Mantel-Haenszel method (Mantel and Haenszel, 1959), a Poisson-normal hierarchical model (Böhning, Mylona, and Kimber, 2015), and a beta-binomial model (Kuss, 2015), among others.

In a logistic regression, if one covariate perfectly predicts the response, the maximum likelihood estimate (MLE) of the corresponding regression coefficient does not exist (Albert and Anderson, 1984). This so-called separation problem occurs in a meta-analysis if all trials include zero events for the same treatment arm. This can be seen as the most extreme example of data sparsity in a meta-analysis of few studies involving rare events. One way to deal with the separation problem in logistic regression is penalization (Greenland and Mansournia, 2015). This is achieved by adding a penalty term to the likelihood in a frequentist framework to penalize estimates of regression coefficients, for example Firth's penalization (Firth, 1993). In a Bayesian framework, one can use a weakly informative prior (WIP) for the same purpose (Gelman, Jakulin, et al., 2008).

The parameter which controls the heterogeneity between trials in a random-effect meta-analysis model is hard to estimate, if the number of studies is low. This is because the heterogeneity parameter is informed by the number of studies which are meta-analyzed. To solve this, the use of WIPs for the heterogeneity parameter has been proposed (Gelman, 2006; Friede et al., 2017). Here, I consider a doubly challenging problem, that is meta-analysis of few studies involving rare events. Inspired by the penalization ideas, I propose to use WIPs both for the treatment effect parameter and the heterogeneity parameter. To my knowledge, our proposal is the only method suggested specifically for meta-analyses of few studies involving rare events. As a data model, I choose the binomial-normal hierarchical model (BNHM) originally introduced by Smith, Spiegelhalter, and A. Thomas (1995). The baseline risks are treated as fixed-effects and relative treatment effects are modeled in this BNHM. I construct a WIP for the treatment effect parameter by assuming priori the expected range for the odds ratio values. Moreover, I re-analyze a large set of meta-analysis datasets from the CDSR to empirically investigate a plausible default WIP for the treatment effect parameter. The proposed method is assessed in a simulation study in terms of some performance measures, namely the mean bias of the treatment effect parameter, the mean coverage probability and the mean length of the interval estimates. To illustrate the proposed methods, I consider a meta-analysis in paediatric liver transplantation.

1.5 Outline

I consider the aforementioned research questions introduced in Chapter 1.4. My research on these questions are published (Günhan, Röver, and Friede, 2020; Günhan, Weber, and Friede, 2020) or accepted for publication (Günhan, Meyvisch, and Friede, 2020) or currently under review (Günhan, Weber, Seroutou, et al., 2020) in peer reviewed journals. I will summarize my investigations in Chapter 2. Chapter 2 includes three sections, each one dealing with one of the research questions introduced in Chapter 1.1. Lastly, I will critically discuss my findings in Chapter 3 and give some thoughts for future research.

2 Proposed Bayesian methods for clinical drug development

2.1 Phase I dose-escalation trials with multiple schedules

In the following, I will summarize my proposed method TITE-PK for phase I dose-escalation trials with multiple schedules. Two types of designs, simultaneous and sequential investigations of multiple schedules are published (Günhan, Weber, and Friede, 2020) or under review (Günhan, Weber, Seroutou, et al., 2020).

A time-varying Poisson process is utilized to model time-to-first dose limiting toxicities (DLT). Hence, the hazard function $h(t)$ is given by

$$h(t) = \beta\, E(t) \tag{2.1}$$

where $E(t)$ is the exposure measure of the drug and β is the only parameter in the model.

The exposure measure $E(t)$ is calculated using a pseudo-PK model. The pseudo-PK model consists of two models, (1) a central compartment model and (2) an effect compartment model (Kallen, 2007, Chapter 2). The former is characterized by

$$\frac{\mathrm{d}C(t)}{\mathrm{d}t} = -k_e\, C(t)$$

where k_e is the elimination rate constant and $C(t)$ is the drug concentration in the central compartment. Then, the effect compartment model is used to take into account the potential delay between the concentration in the central compartment and the concentration during the pharmacodynamic effect:

$$\frac{\mathrm{d}C_{\text{eff}}(t)}{\mathrm{d}t} = k_{\text{eff}}\,(C(t) - C_{\text{eff}}(t))$$

where k_{eff} is the PK parameter which governs the delay and $C_{\text{eff}}(t)$ is the drug concentration in the effect compartment. The exposure measure is assumed to be equal to the drug concentration in the effect compartment, that is $E(t) = C_{\text{eff}}(t)$. The model is conditioned on the PK parameters k_e and k_{eff}, meaning that the PK parameters are assumed to be known.

If both sides of equation (2.1) are integrated, the cumulative hazard function $H(t)$ is obtained, i. e.

$$H(t) = \beta\, \mathrm{AUC}_E(t) \tag{2.2}$$

where $\mathrm{AUC}_E(t)$ is the area under the curve of the exposure measure. In order to write the likelihood function by taking into account the censored patients, the probability density function $f(t)$ and the survivor function $S(t)$ are required. From the

fundamental relationships of survival analysis, $f(t)$ and $S(t)$ are given by (Kalbfleisch and Prentice, 2002)

$$\begin{aligned} f(t) &= h(t)\ \exp(-H(t)) \text{and} \\ S(t) &= P(T > t) = \exp(-H(t)). \end{aligned} \tag{2.3}$$

Patients with DLT are censored at the time of DLT. The remaining patients are censored at the end of cycle 1, that is $C_j = t^*$ where C_j refers to the censoring time for patient j and t^* denotes the end of cycle 1. Let δ_j be an event indicator, which is 0 for censored events and 1 for DLT events. The likelihood function then can be written as

$$L(T, C|\beta) = \prod_{j=1}^{J} f(T_j|\beta)^{\delta_j}\, S(C_j|\beta)^{(1-\delta_j)}$$

where J is the total number of the patients.

As a measure for the dose-schedule decisions, the probability that a patient experiences a DLT within the first cycle given the dose d and the frequency of administration f, $P(T \leq t^*|d, f)$, or, in short, the end-of-cycle 1 DLT probability is used. This is similar to the use of DLT probabilities in a standard method like the Continual Reassessment Method (CRM) (O'Quigley, Pepe, and Fisher, 1990). From equation (2.3), it follows that

$$\mathrm{P}(T \leq t^*|d, f) = 1 - \exp(-H(t^*|d, f)). \tag{2.4}$$

TITE-PK uses an adapted escalation with overdose control (EWOC) (Babb, Rogatko, and Zacks, 1998) criterion in order to inform dose-schedule decisions. For this purpose, the end-of-cycle 1 DLT probabilities $P(T \leq t^*|d, f)$ are divided into three categories:

(i) $\mathrm{P}(T \leq t^*|d, f) < \alpha_1$ Underdosing (UD)

(ii) $\alpha_1 \leq \mathrm{P}(T \leq t^*|d, f) \leq \alpha_2$ Targeted toxicity (TT)

(iii) $\alpha_2 < \mathrm{P}(T \leq t^*|d, f)$ Overdosing (OD)

If the overdosing probability $\mathrm{P}(\mathrm{P}(T \leq t^*|d, f) > \alpha_2)$ of the dose-schedule combination is higher than a pre-specified feasibility bound a, the corresponding dose-schedule combination is regarded as overly toxic based on the EWOC criterion (Babb, Rogatko, and Zacks, 1998).

Since TITE-PK is fitted in a Bayesian framework, the prior specification for the parameter β is crucial. To inform the prior specification, I establish a relationship between the model parameter β and the end-of-cycle 1 probability $P(T \leq t^*|d, f)$. By combining equations (2.2) and (2.4), I obtain the following relationship between the end-of-cycle 1 probability and the parameter β:

$$\mathrm{cloglog}(\mathrm{P}(T \leq t^*|d, f)) = \log(\beta) + \log(\mathrm{AUC}_E(t^*|d, f)) \tag{2.5}$$

where $\mathrm{cloglog}(x) = \log(-\log(1 - x))$.

The exposure measure $E(t|d, f)$ is re-scaled using a reference dose d^* and a reference frequency of administration f^* such that

$$\int_0^{t^*} E(t|d^*, f^*)\,\mathrm{d}t = \mathrm{AUC}_E(t^*|d^*, f^*) = 1.$$

For the reference dose d^* and the reference frequency of administration f^*, equation (2.5) becomes $\mathrm{cloglog}(P(T \leq t^*|d^*, f^*)) = \log(\beta)$. Using this relationship, the prior in TITE-PK is specified on the end-of-cycle 1 probability which has an easier interpretation compared to the parameter β.

In standard methods like the CRM, the underlying assumption is the monotonicity of the DLT probability in the dose. Similar to this assumption, TITE-PK assumes the monotonicity of the end-of-cycle 1 DLT probability in the exposure measure.

2.1.1 Simultaneous investigation of multiple schedules

In this section, I will summarize the TITE-PK method for the simultaneous investigation of multiple schedules in a phase I dose-escalation trial, which was published in Günhan, Weber, and Friede (2020). As described in Chapter 1.4.1, both doses and schedules are allowed to vary in the simultaneous investigation of multiple schedules. TITE-PK is able to account for multiple schedules using a pseudo-PK model. In the trial, the dose-schedule combination for the next cohort is chosen as the combination with the lowest $\mathrm{AUC}_E(t^*)$ among the combinations which are not overly toxic based on the EWOC criterion.

To assess the long-run properties of the TITE-PK method in a simultaneous design and compare it to the POCRM (Wages, O'Quigley, and Conaway, 2014), I conduct a simulation study. The simulation settings are inspired by the Vidaza trial. Three doses (8, 16, and 24 $\mathrm{mg/m^2}$) and four schedules (A, B, C, and D) are investigated in the simulations as in the Vidaza trial. In the Vidaza trial, schedules refer to the number of treatment cycles included in the trial. However, I use another definition of the schedule in my research, that is the frequency of administration. This definition is seen as more practical by Guo, Li, and Yuan (2016), among others. For this reason, I assume that schedules A, B, C, and D refer to the dosing frequencies of 192, 96, 48, and 24 hours in the simulations, respectively.

In the simulations, different curves of dose-DLT probabilities are investigated. They are displayed in Figure 2.1. In Scenario 1, there is no dose-schedule combination in the overdosing interval. In contrast, all combinations are in the overdosing interval in Scenario 2. Scenarios 3-5 include curves of dose-DLT probabilities which are spread across underdosing, targeted toxicity and overdosing intervals. In Scenarios 1-5, the underlying assumption is that DLT probabilities are monotonically increasing with the more frequently administered schedule, in other words the schedules are ordered completely. Scenario 6 is similar to Scenario 1 with the difference that the DLT probabilities of Schedules B and C are switched. Hence, the assumption of the completely ordered schedules is violated in Scenario 6.

To implement TITE-PK, the PK parameters k_e and k_{eff} must be specified in the pseudo-PK model. It is reported that the elimination half-life of Vidaza is 4 hours, and its absorption is rapid (Celgene, 2018). Consequently, I specify $k_e = \frac{\log(2)}{4}$

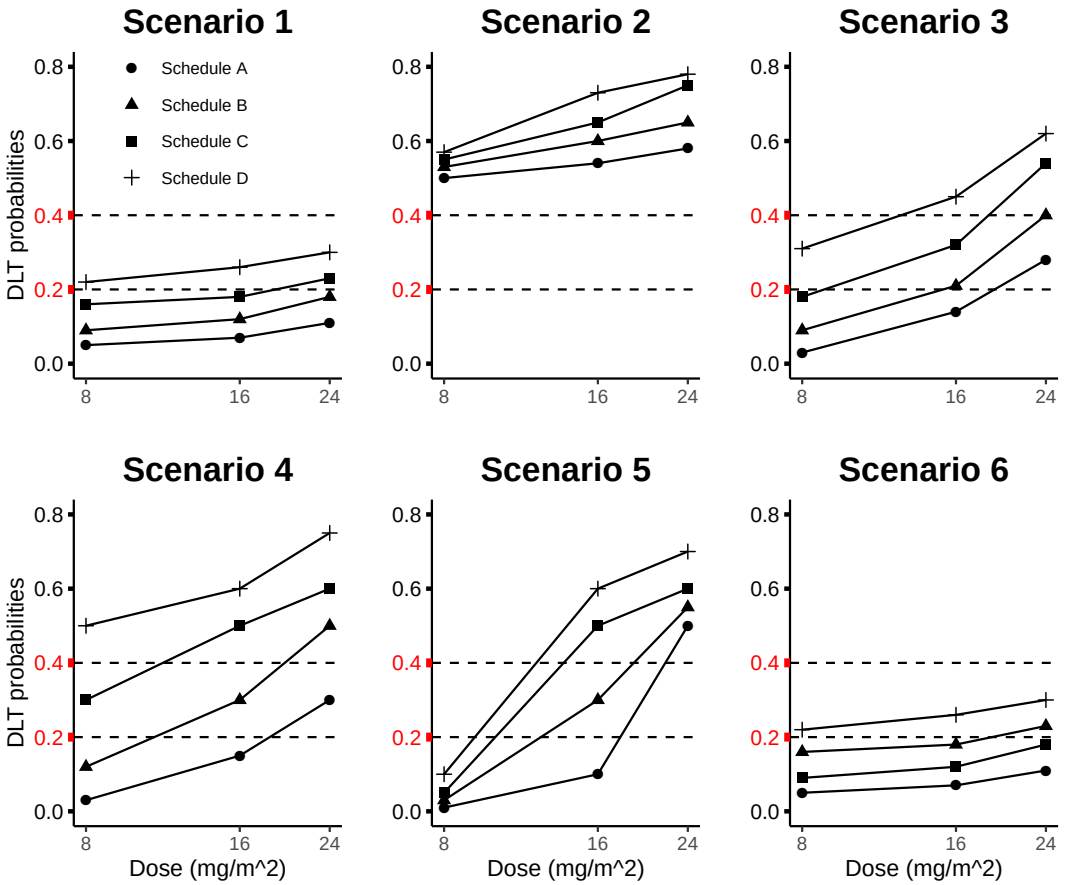

FIGURE 2.1: Simulation scenarios: Each scenario consists of four dose-DLT probability curves for each schedule. Targeted toxicity intervals (0.2 - 0.4) are shown by the horizontal dashed lines. This interval is used to categorize the DLT probabilities of the simulation scenarios. Schedules A, B, C, and D correspond to the frequency of administration of 192, 96, 48, and 24 hours, respectively.

(1/hours). The parameter k_{eff} is determined by assuming a log-normal distribution with the 0.025 and 0.975 quantiles of the absorption rate and the cycle length, resulting in $k_{\text{eff}} = 0.295$. As the prior distribution for the end-of-cycle 1 probability $P(T \leq t^*|d^*, f^*)$, a normal distribution with mean 0.30 and standard deviation 1.75 is used. This prior distribution suggests that a priori the end-of-cycle 1 probability of the reference dose and the reference frequency of administration is 0.30. As the comparator, I used the POCRM method. For both POCRM and TITE-PK, cohorts of size 1 are assumed.

In the following, I will explain the decision criteria used in the simulations for TITE-PK. Before the MTC declaration, a minimum of 9 patients should be treated at the declared MTC. Also, one of two following conditions must hold (1) the probability of the targeted toxicity of the declared MTC must be higher than 50% or (2) at least 21 patients must be treated in the trial. The trial is stopped without any MTC declaration, if the EWOC criterion is not met for any combination. The maximum sample size is set to 60. After exhausting 60 patients, the MTC must be declared or

the trial has to be stopped without any MTC declaration. For TITE-PK implementation, $\alpha_1 = 0.16$, $\alpha_2 = 0.33$, and $a = 0.50$ are used as boundaries of the targeted toxicity interval and the feasibility bound, respectively. Here, the boundaries of the targeted toxicity interval are different than the values used to categorize DLT probabilities of the simulation scenarios. Thus, I use more conservative EWOC criteria for dose-schedule decisions. See Wages, O'Quigley, and Conaway (2014) for the criteria of the dose-schedule decisions of POCRM. 1,000 replications are generated for each scenario.

TABLE 2.1: Six metrics to assess the simulations results obtained by TITE-PK and POCRM. The values of MTC_U, MTC_T, MTC_O, and Stopped are sum up to 100.

Metric	**Definition**
MTC_U	Percentage of trials with MTC in the underdosing interval.
MTC_TT	Percentage of trials with MTC in the targeted toxicity interval
MTC_O	Percentage of trials with MTC in the overdosing interval.
Stopped	Percentage of trials which are stopped without declaring MTC.
Mean_DLT	Mean number of DLT occurred.
Mean_N	Mean number of patients in the trial.

Six popular metrics are used to assess the simulation results obtained by TITE-PK and POCRM, which are listed in Table 2.1. Higher values of MTC_TT mean a higher accuracy of the method, while higher values of MTC_O result in more patients administered with more toxic combinations. Therefore, higher values for MTC_TT and lower values for MTC_O, Mean_N, and Mean_DLT are desirable.

The results obtained by TITE-PK and POCRM are displayed in Figure 2.2. In Scenario 1, there is no dose-schedule combination in the overdosing interval. TITE-PK (0.76) outperforms POCRM (0.65) in terms of the MTC declared in the targeted toxicity interval. In Scenario 2, all combinations are in the overdosing interval. Hence, higher values for the percentage of trials which are stopped without MTC declaration are desirable. TITE-PK results in superior performance in terms of the metric "Stopped" in comparison to POCRM (TITE-PK 0.72 vs POCRM 0.50). In terms of the MTC recommendation in the targeted toxicity interval, TITE-PK produces slightly better results than POCRM in Scenarios 3 and 6. In contrast, POCRM yields slightly better results in Scenarios 4 and 5 in terms of the MTC recommendation in the targeted toxicity interval.

In all scenarios, TITE-PK yields lower percentages compared to POCRM in terms of the MTC recommendation in the overdosing interval, that is the metric MTC_O. TITE-PK produces lower mean number of DLT in comparison to POCRM. This is more pronounced in Scenario 2, where the corresponding values for TITE-PK and POCRM are 8.7 and 17.6, respectively. Furthermore, TITE-PK requires lower mean number of total patients compared to POCRM. In Scenario 1, the corresponding values are 18.8 and 25.9 for TITE-PK and POCRM, respectively.

In summary, the proposed method TITE-PK can be used for phase I dose-escalation trials with simultaneous design of multiple schedules. The dose-schedule decisions are based on the EWOC criterion, which controls the number of patients who are

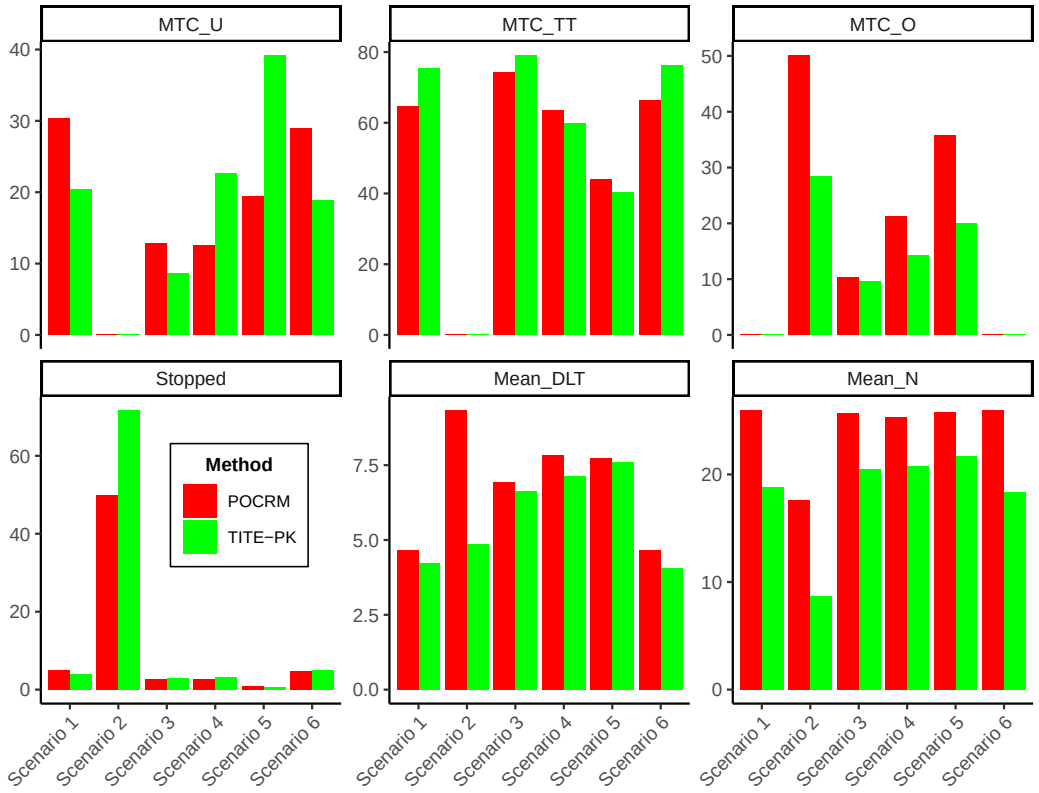

FIGURE 2.2: Simulation results of six scenarios obtained by TITE-PK and POCRM. For each scenario, six metrics are shown as bar plots. Definition of the metrics are given in Table 2.1 and scenarios are demonstrated in Figure 2.1.

exposed to overly toxic dose-schedule combinations. The simulation study demonstrated that TITE-PK outperforms the comparator POCRM in terms of the MTC recommendation in the targeted toxicity and the overdosing intervals in most of the investigated scenarios. **Stan** is used to implement the proposed method TITE-PK. The R and **Stan** code to run the described simulations are publicly available from my personal Github account (`https://github.com/gunhanb/TITEPK_code`). One can adapt the R and **Stan** code based on the application to implement the TITE-PK method.

2.1.2 Sequential investigation of multiple schedules

In this chapter, I will review another application of the TITE-PK method, that is the sequential investigation of multiple schedules in phase I dose-escalation trials. The corresponding manuscript is currently under review (Günhan, Weber, Seroutou, et al., 2020). As explained in Chapter 1.4.1, I assume that there is an ongoing trial and a completed trial with different treatment schedules. The main aim is using the data from both the completed and the ongoing trial to inform dose-escalation decisions in the ongoing trial. For this purpose, TITE-PK is a natural solution, since it is able to integrate multiple schedules. More specifically, after the MTD declaration

TABLE 2.2: The everolimus trial: For each dose-schedule combination, the corresponding doses, treatment schedules, the sample sizes and the number of DLT occurred are listed.

Dose (mg/m^2)	Schedule	Sample size	Number of DLT
20.0	Weekly	5	0
30.0	Weekly	13	4
2.5	Daily	4	2
5.0	Daily	6	3

of the stage with the first schedule, the ongoing trial with another schedule can be informed using both information from the completed and the ongoing trial with TITE-PK.

I used a phase I dose-escalation trial of everolimus as an illustrative application for the use of TITE-PK in the sequential investigation of multiple schedules (Besse et al., 2014). Everolimus was administered to patients together with the standard of care to find a suitable dose and schedule in the treatment of small lung cancer (ClinicalTrials.gov identifier: NCT00466466). In the everolimus trial, daily and weekly schedules were investigated. The final dataset is listed in Table 2.2. All DLT occurred on the 15th day. The length of one cycle of the treatment is 21 days. Doses in two schedules were escalated separately and no information from the other schedule was used to inform dose-escalation decisions. The elimination half-life of everolimus is 30 h and the absorption rate is 2.5 1/h (Novartis, 2012).

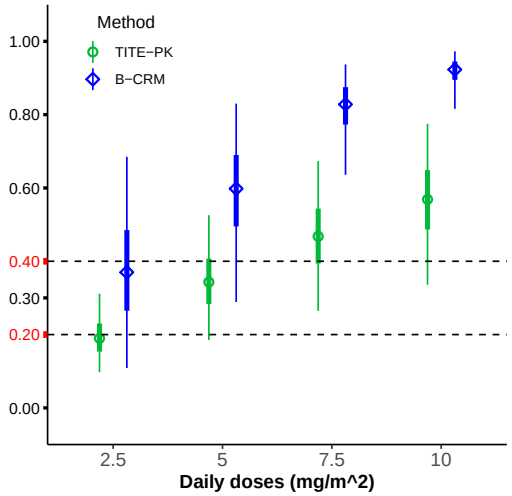

FIGURE 2.3: The everolimus trial: The posterior distribution of DLT probabilities of daily doses obtained by TITE-PK and B-CRM approaches. Horizontal dashed lines [0.20, 0.40] are the boundaries of the targeted toxicity interval. For each DLT probability, the median, the 50%, and the 95% credible intervals are investigated.

To illustrate the proposed TITE-PK method, I analyze the everolimus trial as if it was conducted sequentially. More precisely, I assume that first the weekly schedule

was investigated and then the daily schedule. The idea is to borrow from the data with the weekly schedule when estimating DLT probabilities for the daily schedule. The prior distributions must be specified for the parameters of the bridging CRM (B-CRM) and TITE-PK methods. For TITE-PK, I specify a normal WIP for the end-of-cycle 1 probability, namely $\mathcal{N}(0.30, 1.25^2)$. As the reference combination, I use 5 mg/m^2 as the reference dose and 24 hours (daily schedule) as the reference frequency. Based on the elimination half-life and the absorption rate, the PK parameters are specified as follows: $k_e = \frac{\log(2)}{30}$ and $k_{\text{eff}} = 1.45$. As the boundaries of the targeted toxicity interval (α_1 and α_1), [0.20, 0.40] are used. Also, 0.25 is used for the feasibility bound. The B-CRM is based on a one-parameter power model. A normal prior with mean 0 and standard deviation 2 is used as the prior for the power parameter α in B-CRM. The target probability is 0.30 for B-CRM. The following stopping rule for B-CRM is used: $P(\pi_1 > 0.30) < 0.90$ where π_1 is the DLT probability of the lowest dose (Liu et al., 2015).

In Figure 2.3, the summaries of the posterior distributions of DLT probabilities for each daily dose obtained by B-CRM and TITE-PK are shown. The horizontal dashed lines correspond to the boundaries of the targeted toxicity interval. The points, thick lines, and thin lines refer to the medians, 50% and 95% credible intervals, respectively. For all four doses, B-CRM produces higher DLT probabilities than TITE-PK. For TITE-PK, only 2.5 mg/m^2 is eligible based on the EWOC criterion ($P(P(T \leq t^*|d = 2.5, f = 24) > 0.40) = 0$). Also, B-CRM concludes that 2.5 mg/m^2 is not an overly toxic dose based on its stopping criterion ($P(\pi_1 > 0.30) = 0.67$). Thus, both methods suggest that 2.5 mg/m^2 can be declared as the MTD, confirming the conclusion of the original everolimus trial.

In the following, I will summarize the simulation study conducted to assess the performance of TITE-PK in the sequential investigation of multiple schedules. The settings are inspired by the everolimus trial. Doses of 2.5, 5, 7.5, 10, 12.5, and 15 mg/m^2 are investigated in two schedules, first a schedule of 48 h dosing intervals and then a schedule of 24 h intervals. The investigated scenarios are displayed in Figure 2.4. In Scenario 1, most of the dose-schedule combinations are in the underdosing interval, whereas all combinations are in the overdosing interval in Scenario 5. Scenarios 2-4 are more spread across underdosing, targeted toxicity, and overdosing intervals. Scenario 6 violates the monotonicity assumption between the exposure and the DLT probabilities. This means that given the same dose, the more frequent schedule has lower DLT probability. 1,000 replications are generated for each scenario.

The results obtained by TITE-PK and B-CRM are displayed in Figure 2.5. The metrics from Chapter 2.1.1 are used to assess the performance of the methods, see Table 2.1 for the definitions. In Scenarios 1-4, TITE-PK outperforms B-CRM in terms of the MTD declaration in the targeted toxicity interval. TITE-PK achieves this by selecting the MTD in the targeted toxicity interval in 7%, 20%, 30%, 24%, and 10% more simulated trials. In Scenario 2, TITE-PK yields lower percentage in terms of the MTD declaration in the overdosing interval compared to B-CRM (38% vs 22%). In Scenario 5, TITE-PK stops the trial in 98% of the time, while B-CRM only stops the trial in 75% of the time. In Scenario 6, B-CRM clearly outperforms TITE-PK in terms of the MTD declaration in the targeted toxicity interval (77% vs 17%). Recall

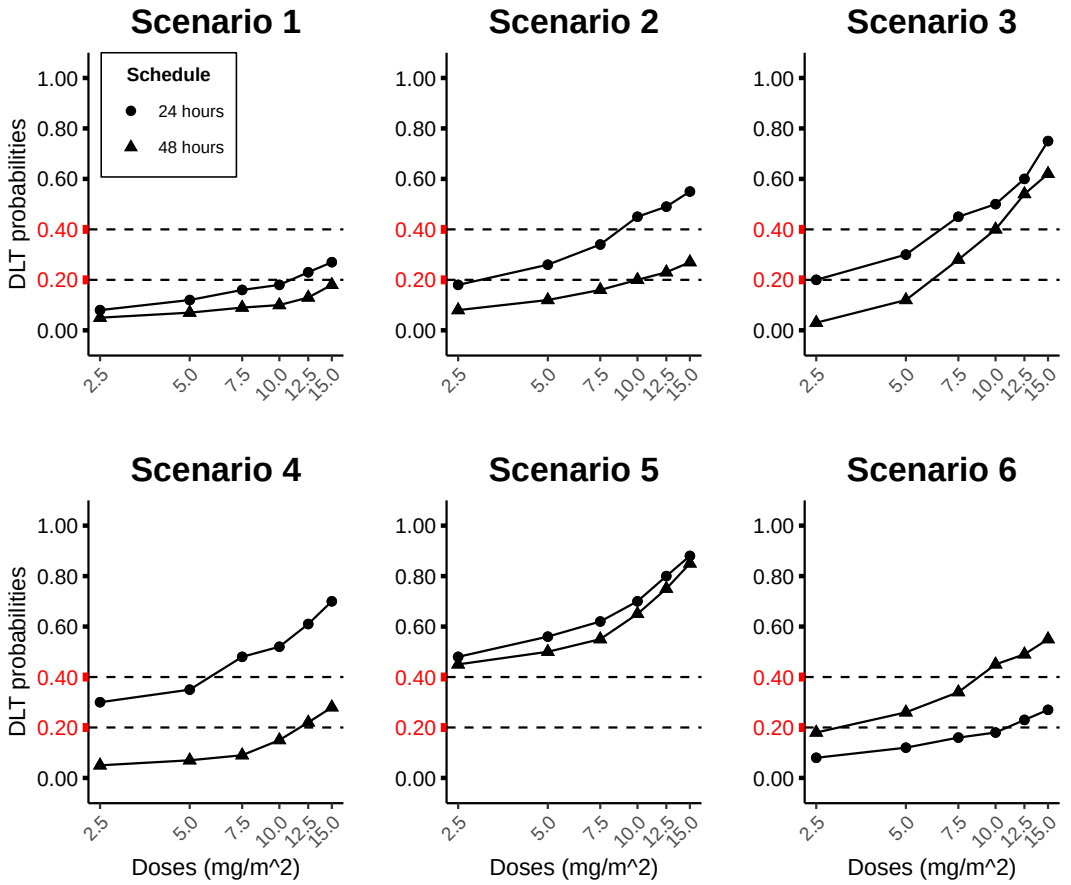

FIGURE 2.4: Simulation scenarios: Each scenario consists of two dose-DLT probability curves for the weekly and the daily schedule. Weekly dose levels are 20, 30, 40, 50, 60, and 70 mg/m^2 and daily dose levels are 2.5, 5, 7.5, 10, 12.5, and 15 mg/m^2. Targeted toxicity intervals (0.20 - 0.40) are shown by the horizontal dashed lines.

that in Scenario 6, the monotonicity assumption between the exposure and the DLT probabilities is violated. In half of the scenarios, the mean number of DLT occurred in the trial is lower in TITE-PK compared to B-CRM. In Scenarios 1-4 and 6, the mean number of sample sizes are slightly higher in TITE-PK in comparison to B-CRM.

In conclusion, the proposed method TITE-PK can be used as an alternative to the B-CRM approach to design and analyze a sequential phase I dose-escalation trial with multiple schedules. A phase I trial involving daily and weekly schedules was used to illustrate the use of TITE-PK. In simulations, TITE-PK displays better performance than the comparator B-CRM in most of scenarios in terms of the declaring MTD in the targeted toxicity interval. These results come with slightly larger mean number of DLT and sample sizes, while the MTD declaration is less often in the overdosing interval (except for Scenario 4), thus safer for the patients in the long term compared to the B-CRM. However, when the monotonicity assumption between exposure and DLT probabilities does not hold (Scenario 6), B-CRM yields better results. Hence, I suggest the use of the TITE-PK method, if a heavy violation of the monotonicity assumption is not expected. The **Stan** and R code

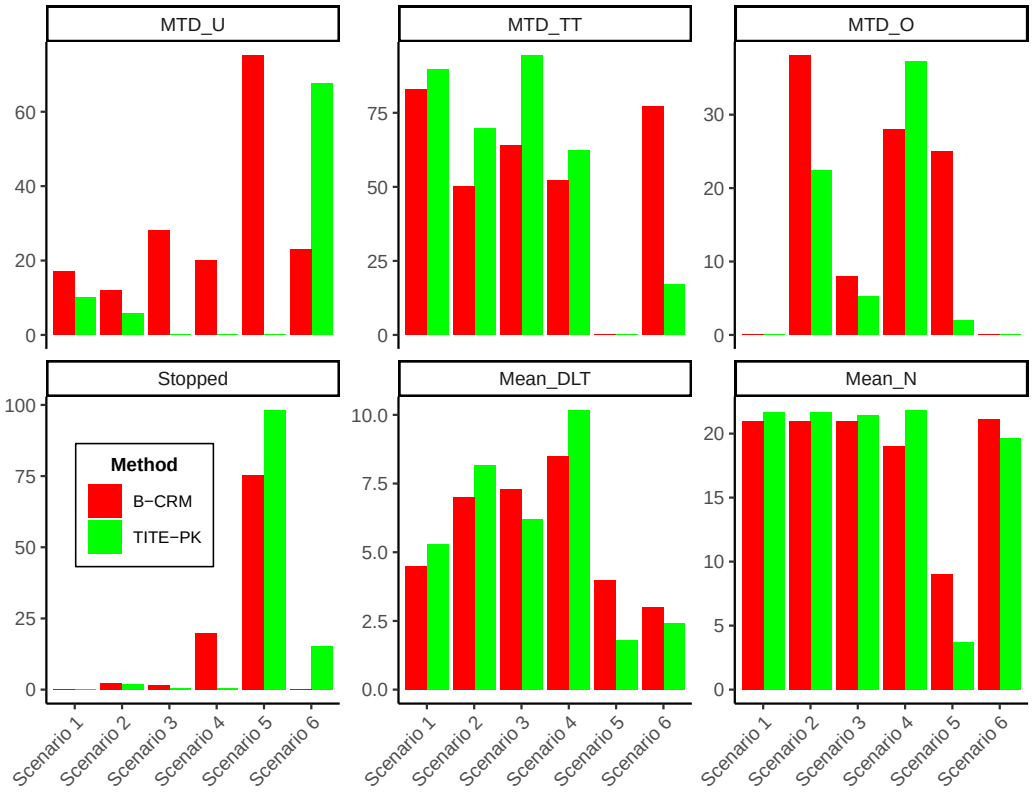

FIGURE 2.5: Simulation results of six scenarios obtained by TITE-PK and B-CRM. For each scenario, six metrics are shown as bar plots. Definition of the metrics are given in Table 2.1 and scenarios are demonstrated in Figure 2.4.

to analyze the everolimus trial using TITE-PK is publicly available from Github (`https://github.com/gunhanb/TITEPK_sequential`), facilitating the application of the proposed method.

2.2 Phase II dose-finding trials with multiple schedules

In this chapter, I will briefly describe my research on phase II dose-finding trials with multiple schedules, which is accepted for publication (Günhan, Meyvisch, and Friede, 2020). I propose the use of schedule-specific random effects in a partial pooling approach to model heterogeneity in the model parameters between schedules in a phase II dose-finding trial. I will review a simulation study to compare the proposed method to several alternatives. A phase II trial in atopic dermatitis is re-analyzed to illustrate the proposed method.

For the data model, I assume that the response y_{ijk} for schedule i, dose j, and patient k is normally distributed (Feller et al., 2017), i.e.

$$y_{ijk} \sim \mathcal{N}(f(d_j^{(i)}, \boldsymbol{\theta}), \sigma_i^2)$$

where $\boldsymbol{\theta}$ is the vector of the model parameters ($\mathrm{E}_0^{(i)}$, $\mathrm{E}_{\max}^{(i)}$ and $\mathrm{ED}_{50}^{(i)}$) and σ_i is the standard deviation of the error terms. The $f(d_j^{(i)})$ characterizes the relationship between the dose and the response for each schedule. For $f(d_j^{(i)})$, I use the popular Emax model (N. Thomas, Sweeney, and Somayaji, 2014), i.e.

$$f(d_j^{(i)}, \boldsymbol{\theta}) = \mathrm{E}_0^{(i)} + \mathrm{E}_{\max}^{(i)} \frac{d_j^{(i)}}{\mathrm{ED}_{50}^{(i)} + d_j^{(i)}}$$

where $\mathrm{E}_0^{(i)}$ and $\mathrm{E}_{\max}^{(i)}$ represent the placebo effect and the maximal effect attributable to the drug, respectively. $\mathrm{ED}_{50}^{(i)}$ is the dose providing 50% of the maximal effect.

My aim is estimating dose-response curves for each schedule. One way is to assume that all parameters are shared between different schedules, hence conducting a complete pooling approach. This means that assuming $\mathrm{E}_0^{(1)} = \mathrm{E}_0^{(2)} = \ldots$ and $\mathrm{E}_{\max}^{(1)} = \mathrm{E}_{\max}^{(2)} = \ldots$. For $\mathrm{ED}_{50}^{(i)}$, I use the re-scaled $\mathrm{ED}_{50}^{(i)}$ parameters. For this reason, I specify a *reference schedule* (i_{ref}). The re-scaled parameters are given by $\mathrm{ED}_{50}^{*(i)} = \mathrm{ED}_{50}^{(i)} \frac{f^{(i)}}{f^{(i_{\mathrm{ref}})}}$ where $f^{(i_{\mathrm{ref}})}$ and $f^{(i)}$ are the frequency of administration of the reference schedule i_{ref} and the schedule i, respectively. However, complete pooling ignores the potential heterogeneity in parameters between schedules.

Alternatively, Feller et al. (2017) suggested the use of schedule specific fixed-effects for the parameters $\mathrm{E}_{\max}^{(i)}$ and/or $\mathrm{ED}_{50}^{(i)}$, while assuming a common parameter for the $\mathrm{E}_0^{(i)}$. In other words, this is partial pooling with fixed-effects (PP - FE). In this method, the rescaling of $\mathrm{ED}_{50}^{(i)}$ is not needed, since fixed-effect parameters $\mathrm{ED}_{50}^{(i)}$ for each schedule are estimated. Instead of the schedule specific fixed-effects like in PP - FE, I suggest the use of schedule specific random-effects for the parameters $\mathrm{E}_{\max}^{(i)}$ and/or $\mathrm{ED}_{50}^{(i)}$, while assuming a common parameter for the $\mathrm{E}_0^{(i)}$. Hence, my proposed model is partial pooling with random-effects (PP - RE). The schedule specific random-effects can be obtained by assuming exchangeable random-effects around an overall mean. Different $\mathrm{ED}_{50}^{(i)}$ parameters are transformed into the same scale by using a reference schedule. The re-scaled parameters can be obtained by

$\text{ED}_{50}^{*(i)} = \text{ED}_{50}^{(i)} \left(f^{(i)} / f^{(i_{\text{ref}})}\right)$ as in the complete pooling. Using the log transformation, I assume that

$$\log(\text{ED}_{50}^{*(i)}) \sim \mathcal{N}(\mu_{\text{ED}_{50}}, \tau^2_{\text{ED}_{50}}) \tag{2.6}$$

where $\mu_{\text{ED}_{50}}$ and $\tau_{\text{ED}_{50}}$ refer to the overall mean and the heterogeneity in $\log(\text{ED}_{50}^{*(i)})$ between schedules. To model schedule specific random-effects of the $\text{E}_{\text{max}}^{(i)}$, a re-scaling or the log transformation is not required.

I implement the proposed model in a Bayesian framework. Hence, the choice of prior distributions is very crucial. Noninformative priors are used for the parameters $\text{E}_0^{(i)}$ and $\text{E}_{\text{max}}^{(i)}$, namely $\mathcal{N}(0, 100^2)$, and also for σ_i, that is a half-normal prior with scale 100 $\mathcal{HN}(100)$. It has been shown that the prior choice for $\text{ED}_{50}^{(i)}$ can have significant influence on the posterior distributions (Bornkamp, 2014). As the prior of $\text{ED}_{50}^{(i)}$, I use the functional uniform prior, developed by Bornkamp (2012). Instead of putting a uniform prior directly on the $\text{ED}_{50}^{(i)}$, the functional uniform prior assumes uniformity on the different shapes of the underlying Emax model. The heterogeneity parameter is mainly informed by the number of the schedules in the trial, which is typically low. Hence, this problem is similar to estimating the heterogeneity in treatment effects between trials in the meta-analysis of few studies (Friede et al., 2017). Therefore, for the heterogeneity parameter $\tau_{\text{ED}_{50}}$, I use a weakly informative prior (WIP), namely $\mathcal{HN}(1)$, following the suggestions for a meta-analysis model, when log odds ratio (or log hazard ratio) is used as the effect measure (Friede et al., 2017).

TABLE 2.3: Dupilumab trial: Schedule, dose per administration, and sample size for each arm.

Arm	Schedule	Dose (mg/m^2)	Sample size
1	Weekly	0	61
2	Weekly	300	63
3	Biweekly	200	61
4	Biweekly	300	64
5	Monthly	100	65
6	Monthly	300	65

I consider a phase II trial of dupilumab in atopic dermatitis to illustrate the proposed method (ClinicalTrials.gov identifier: NCT01859988). The dupilumab trial involves three schedules, weekly, biweekly, and monthly schedules. The design of the trial is given in Table 2.3. The primary endpoint of the trial is the percentage change from baseline in Eczema Area and Severity Index (EASI) score at the 85th day. The EASI has values between 0 and 72, which indicates the severity of the eczema. The higher EASI score means higher severity in eczema. The dataset is reported as the least square means and the standard errors for each arm by Thaçi et al. (2016), which is displayed in Figure 2.6A. For each schedule, information from only two arms are available, thus suggesting some data sparsity issue. I analyze the dupilumab trial by using the complete pooling (CP), the partial pooling with fixed-effects (PP - FE), and the partial pooling with random-effects (PP- RE) in a Bayesian framework. For

both partial pooling approaches, the parameters $E_0^{(i)}$, $E_{max}^{(i)}$, and σ_i are assumed to be shared between schedules. The $ED_{50}^{(i)}$ are treated as fixed-effects in PP - FE and as random-effects in PP - RE.

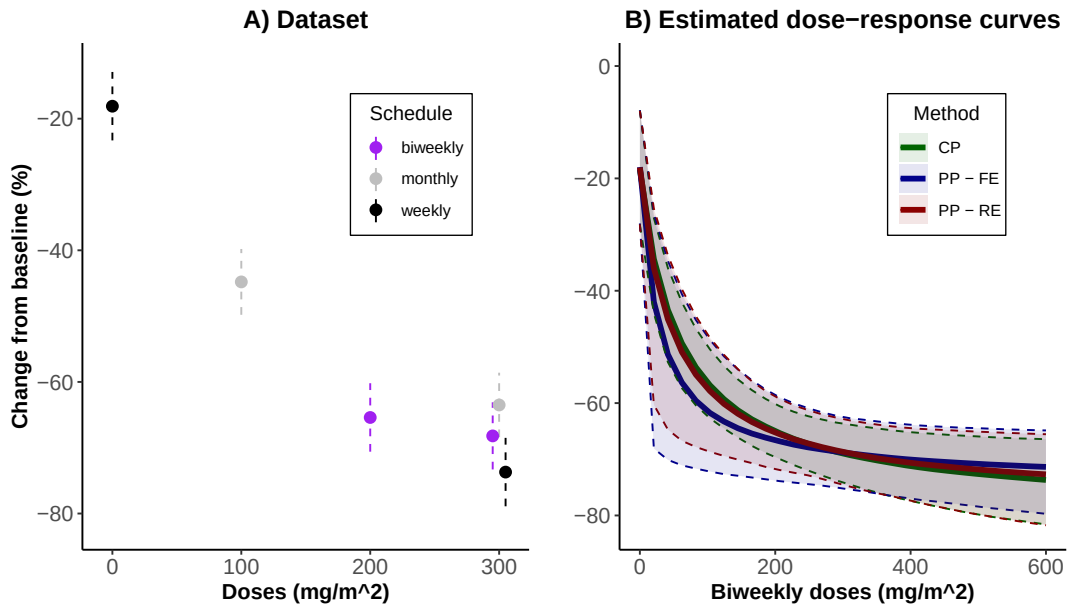

FIGURE 2.6: Dupilumab trial: A) The least square means and 95% confidence intervals of the percentage change from baseline in EASI score at the 85th day for each arm, and B) the estimated dose-response function f obtained by the complete pooling (CP), the partial pooling approach with schedule specific fixed-effects (PP - FE), and the partial pooling approach with schedule specific random-effects (PP - RE) are displayed. Dashed lines in plot B refers to 95% credible intervals of the estimated dose-response function f obtained by three methods.

The estimated dose-response functions of the biweekly dose obtained by three methods are demonstrated in Figure 2.6B. The lines and the colored areas correspond to the posterior medians and the 95% pointwise credible intervals evaluated at a grid. The posterior medians of complete pooling and the partial pooling with random-effects look very similar, while the partial pooling with fixed-effects yields slightly different posterior medians. I also computed the approximate leave-one-out cross-validation information criteria (LOO-IC) (Vehtari, Gelman, and Gabry, 2017) to compare the three models. The corresponding LOO-IC values are 37.0, 37.5, and 39.4 for the complete pooling, the partial pooling with random-effects, and the partial pooling with fixed-effects, respectively. The lower value of LOO-IC indicates better model performance. The reason of the relatively worse performance of the partial pooling with fixed-effects can be explained by data sparsity in the dupilumab trial, hence the simpler models are preferred. Furthermore, the complete pooling displays the shortest credible intervals, that is also observed in the simulations, which I summarize in the following.

To assess the performance of the the proposed method PP - RE and compared to some alternatives, I conducted a simulation study. The simulations are inspired by the dupilumab trial and the MOR106 trial (ClinicalTrials.gov identifier: NCT03568071).

Like the dupilumab trial, the MOR106 trial is a phase II trial in atopic dermatitis, but involves two schedules, biweekly and monthly. In simulations, a trial involves seven arms including a placebo arm, and 1, 3, 10 mg/kg for biweekly and monthly schedules. Note that these doses are doses per administration. As primary outcome, the percentage change in baseline of EASI score is used. As the data-generating process for the dose-response model, an Emax model is used. The values for the parameters $E_0^{(i)} = -20\%$, $E_{max}^{(i)} = -60\%$, $ED_{50}^{biweekly} = 2$ mg/kg, and $\sigma_i = 35\%$ are used. The value for the $ED_{50}^{monthly}$ is varied, namely $ED_{50}^{monthly} \in \{1, 2, 3, 3.5, 4, 4.5, 5, 6, 10$ (mg/kg)$\}$. The scenario of $ED_{50}^{monthly} = 4$ mg/kg corresponds to zero heterogeneity in ED_{50}^{i}, since $2 \times ED_{50}^{weekly} = 4$ mg/kg. 1 000 replications are generated for each scenario.

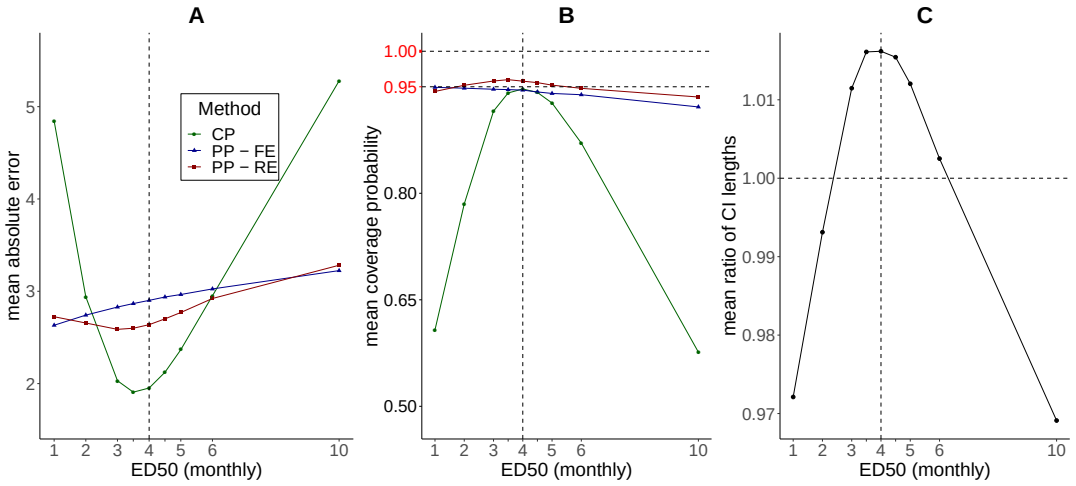

FIGURE 2.7: Simulation results: The mean absolute error (A) and the mean coverage probability (B) for the dose-response function obtained by the complete pooling (CP), schedule specific fixed-effects in a partial pooling approach (PP - FE), and schedule specific random-effects in a partial pooling approach (PP - RE) are displayed. Also, the ratio of the credible interval estimates (C) obtained by the PP - RE and the PP - FE are shown. The denominator of the ratio is the estimates by the PP - RE.

Three methods are compared in a Bayesian framework: the complete pooling, the partial pooling with random-effects, and the partial pooling with fixed-effects. The mean absolute error and the mean coverage probabilities for the three methods are displayed in Figure 2.7A and B. When $ED_{50}^{monthly}$ is close to 4, in other words in case of low heterogeneity, the complete pooling outperforms other methods in terms of the mean absolute error and the mean coverage probability. However, partial pooling approaches yield more robust performance against the change in the $ED_{50}^{monthly}$ compared to the complete pooling. Furthermore, partial pooling with random-effects results in smaller mean absolute errors and higher coverage probabilities compared to the partial pooling with fixed-effects in most of the scenarios. The complete pooling produces the shortest credible intervals in all of the scenarios (results not shown). Figure 2.7C displays the mean ratio of 95% credible intervals by the partial pooling with fixed-effects and the partial pooling with random-effects,

while the latter is the denominator of the ratio. The partial pooling with random-effects yields shorter credible intervals than the partial pooling with fixed-effects in most of the scenarios, hence it is more preferable.

In summary, I propose to use schedule specific random-effects in a partial pooling approach to deal with potential heterogeneity in model parameters for phase II trials with multiple schedules. The proposed method is illustrated with a phase II trial in atopic dermatitis. In simulations, I showed that the use of random-effects exhibits more robust results in terms of the mean absolute error and the coverage probability compared to the complete pooling, while gaining efficiency in comparison to separate analyses of different schedules. Moreover, the proposed method results in smaller mean absolute errors than the partial pooling with fixed-effects in most of the investigated scenarios. I implement the proposed method in an R package, `ModStan`, made publicly available in Github (https://github.com/gunhanb/ModStan).

2.3 Meta-analysis of few studies involving rare events

As described in Chapter 1.4.3, meta-analysis of few studies involving rare events is challenging due to the sparsity of the data. Detailed results of my research on this topic can be found in Günhan, Röver, and Friede (2020), and will be summarized in this section.

The data model has been originally introduced by Smith, Spiegelhalter, and A. Thomas (1995). The number of events r_{ij} are assumed to be binomially distributed $r_{ij} \sim \text{Bin}(\pi_{ij}, n_{ij})$ for each trial i and treatment arm $j \in \{0, 1\}$. Using the logit link, event probabilities π_{ij} are transformed onto the log odds scale

$$\text{logit}(\pi_{ij}) = \mu_i + \theta_i \, x_{ij} \tag{2.7}$$

where x_{ij} is a treatment indicator coded as $+0.5$ for experimental arm (j = 1) and -0.5 = control arm (j = 0). The baseline risks μ_i are treated as trial specific fixed-effects. The relative treatment effects θ_i are assumed to be trial specific random-effects, more specifically $\theta_i \sim \mathcal{N}(\theta, \tau^2)$ where θ is the mean treatment effect and τ is the heterogeneity in treatment effects between trials. The heterogeneity parameter τ gives the information about the degree of the heterogeneity between trials. This model is called the binomial-normal hierarchical model (BNHM).

In a Bayesian framework, the prior distributions for the model parameters (μ_i, θ, and τ) must be specified. A normal distribution with mean 0 and standard deviation 10 is used as the prior for the baseline risks μ_i (Gelman, Jakulin, et al., 2008). The prior choice of the heterogeneity parameter τ has crucial influence on the posterior estimates, especially when the number of studies is small. Following the suggestions given by Friede et al. (2017), I used a WIP for τ, that is a half-normal distribution with scale 0.5, $\mathcal{HN}(0.5)$. The median value of $\mathcal{HN}(0.5)$ is 0.337 with an upper 95% quantile of 0.98, covering plausible values for the heterogeneity parameter τ.

Instead of using a noninformative prior for the treatment effect parameter θ, I propose to use a WIP for θ which is on the log odds ratio scale. The derivation of a WIP for the parameter θ is explained in the following. I assume that the prior distribution is a normal distribution with mean 0 and standard deviation σ_{prior},

$\mathcal{N}(0, \sigma^2_{\text{prior}})$, resulting in equal event probabilities for positive and negative treatment effects. Hence, only the prior standard deviation σ_{prior} needs to be determined to construct the WIP. Assume that a priori the odds ratio $\exp(\theta)$ is within the interval $[1/\delta, \delta]$ with 95% probability where δ is a pre-specified value, which can be written as:

$$P(1/\delta < \exp(\theta) < \delta) = 95\%.$$

Using the 97.5% quantile of the standard normal distribution, the prior standard deviation can be calculated as follows:

$$\sigma_{\text{prior}} = \frac{\log(\delta)}{1.96}. \tag{2.8}$$

Conservatively specifying $\delta = 250$, the corresponding WIP becomes $\mathcal{N}(0, 2.82^2)$. In other words, the prior $\mathcal{N}(0, 2.82^2)$ for θ means that a priori the odds ratio is within the intervals $[1/250, 250]$ with 95% probability.

To empirically investigate a plausible default prior distribution for the treatment effect parameter θ, I re-analyzed meta-analyses from the Cochrane Database of Systematic Review (CDSR). For this purpose, I downloaded all meta-analysis datasets with dichotomous endpoints available on March 2018 from the Cochrane Library website (`https://www.cochranelibrary.com`). This is done using the program `Cochrane_scraper` (Springate, 2018). This procedure results in 37 773 meta-analysis datasets. I re-analyzed the downloaded datasets using the BNHM in a frequentist framework via the `lme4` R package (Bates et al., 2015). Figure 2.8 demonstrates the histogram of the estimates for θ. The 2.5% and 97.5% quantiles of the estimates of θ are -1.94 and 2.06, respectively. These results can be considered as an informal validation of our proposed WIP for θ, that is $\mathcal{N}(0, 2.82^2)$.

The long run properties of the use of WIP for the treatment effect parameter θ in a meta-analysis of few studies involving rare events is assessed in a simulation study. As data-generating process, the BNHM is used. The number of trials included in a meta-analysis is taken as three. The sample size of each trial is generated from the log-normal distribution with mean 5 and standard deviation 1, $\mathcal{LN}(1, 5)$. The parameter values of $\mathcal{LN}(5, 1)$ is obtained by fitting a log-normal distribution to the sample sizes obtained from the CDSR (Kuss, 2015). Note that the median value of $\mathcal{LN}(5, 1)$ is $\exp(5) \approx 148$. Once the sample size of a trial is determined, the sample sizes for control and treatment arms are generated according to a binomial probability of 0.5. Since I focus on the meta-analysis involving rare events, baseline risks on the probability scale are assumed to be in the interval [0.005, 0.05]. To reflect the moderate heterogeneity in treatment effects, the heterogeneity parameter τ is assumed to be 0.28. The value of 0.28 is the median value of the predictive distribution for τ calculated by Turner, Jackson, et al. (2015) using the CDSR. Eleven scenarios are generated by varying the true treatment effect θ, namely $\theta \in \{-5, -4, -3, -2, -1, -0.5, 0, 0.5, 1, 2, 3, 4, 5\}$. For each scenario, 10 000 replications are generated. As mentioned in Section 1.4.3, the data sparsity in a meta-analysis involving rare events is usually reflected in the proportion of the zero studies. Figure 2.9 displays the mean proportion of single- and double-zero studies in a simulated meta-analysis dataset in our simulations. The proportion of the single-

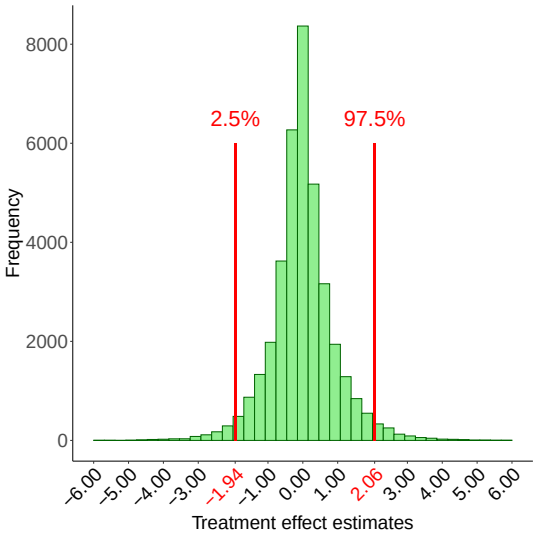

FIGURE 2.8: The distribution of the estimated treatment effect parameter $\hat{\theta}$, obtained from the re-analysis of meta-analysis datasets in Cochrane Database of Systematic Reviews (CDSR). For the analysis, the binomial-normal hierarchical model via maximum likelihood estimations was used.

and double-zero studies decreases with the increase of true value of the treatment effect.

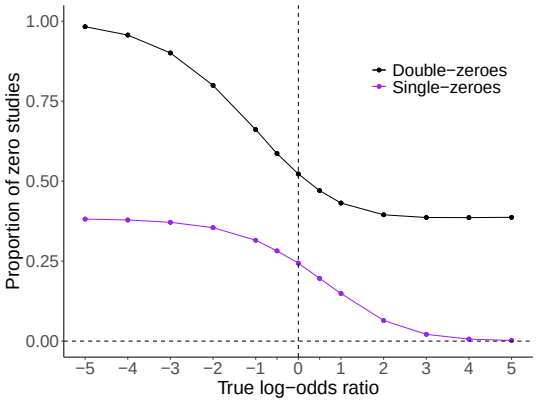

FIGURE 2.9: The mean proportion of zero studies (single- and double-zero studies) in a simulated meta-analysis dataset in our simulations.

In the simulations, I investigated three methods: (1) "WIP", BNHM using a WIP ($\mathcal{N}(0, 2.82^2)$) for θ, (2) "Vague" , BNHM using a vague prior ($\mathcal{N}(0, 100^2)$) for θ and (3) "MLE", BNHM using the MLE. The WIP and Vague are fitted in a Bayesian framework using **Stan**, while the MLE is fitted in a likelihood framework using the `lme4` R package. For the WIP and Vague, a WIP ($\mathcal{HN}(0.5)$) is used for the heterogeneity parameter τ. 2 000 MCMC iterations including 1 000 iterations of burn-in were used with three chains to obtain posterior estimates in the WIP and Vague models. These MCMC settings were tested in some replications in terms of convergence diagnostics, namely $\hat{R}$ and traceplots. I assumed that convergence is reached for

all 10 000 replications for each scenario. For the MLE, the replications in which the convergence is not reached were excluded from the calculation of the performance measures.

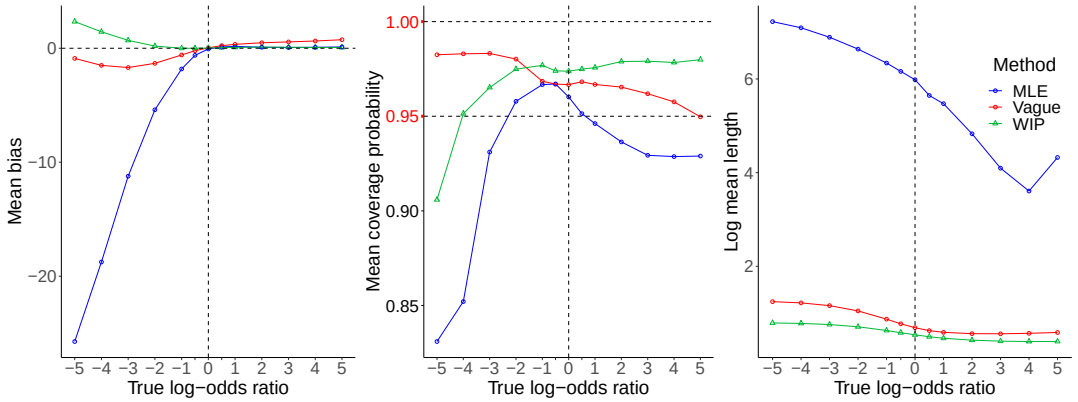

FIGURE 2.10: Simulation results: The mean bias for θ, mean coverage probability of the interval estimates, and the log mean length of the interval estimates for θ obtained by the WIP, Vague, and MLE methods. See the text for the description of the methods.

Three performance measures are used: (1) mean bias, $\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}-\theta)$, (2) mean coverage probability of the interval estimates for θ, and (3) mean length of the interval estimates θ. The mean bias is based on the posterior median for the WIP and Vague approaches, and the maximum likelihood estimate for the MLE. Lower bias, 95% of coverage probability, and shorter interval length are desirable. Simulation results are displayed in Figure 2.10. The WIP outperforms the Vague and the MLE in terms of bias. All three methods improve in terms of bias, when true log odds ratio increases. This is because the data sparsity decreases with the increase of the true log-odds ratio (see Figure 2.9). In terms of the coverage probability of the interval estimates for θ, the Vague and the WIP result in probabilities higher than 0.95, while the MLE shows lower coverage than the nominal value for most of the scenarios. This behaviour of the MLE was also observed by Friede et al. (2017). The WIP displays shorter interval estimates than the Vague and the MLE in all of the scenarios. Overall, the proposed method (WIP) shows superior performance in comparison to the Vague and the MLE in terms of the investigated measures.

To illustrate the proposed method, I re-analyzed a meta-analysis dataset in pediatric liver transplantation (Crins et al., 2014). The dataset includes trials in which the Interleukin-2 receptor antibodies basiliximab and daclizumab are used to decrease the risk of acute rejection of liver transplants. For illustrative purposes, I focus on the outcome post-transplant lymphoproliferative disease (PTLD). The PTLD counts and the sample sizes are given in the top panel of Figure 2.11. The dataset includes three trials including one single-zero trial, one double-zero trial. The results obtained by the WIP, the Vague and the MLE are shown in the bottom panel of Figure 2.11. Point estimates of the three methods are very close. The WIP results in shorter interval estimates compared to the Vague. This behaviour was also observed in the simulations. The MLE yields zero heterogeneity, whereas the WIP and the Vague produce 0.32 and 0.34 (posterior median), respectively. Since, the WIP and the Vague uses same weakly informative prior for τ, similar estimates for τ are expected.

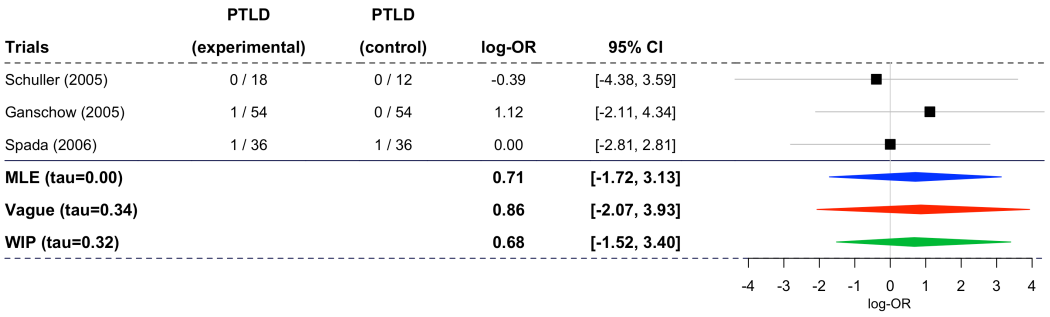

FIGURE 2.11: The post-transplant lymphoproliferative disease (PTLD) counts and sample sizes for three trials are given at the top panel. In the bottom panel, the results obtained by three methods (the WIP, the Vague, and the MLE) are displayed. Moreover, the heterogeneity estimates from the three methods are shown on the left.

The Bayesian version of the BNHM was implemented using **Stan**. To make it accessible for practitioners, I developed an R package, `MetaStan`, which implements the BNHM using **Stan**. `MetaStan` is available from CRAN (https://CRAN.R-project.org/package=MetaStan). More detailed explanations about the package are available from the vignette of the package, which can be obtained running the command `vignette("MetaStan_BNHM")` in an R console.

In summary, I propose to use a WIP for the treatment effect parameter of the BNHM for the meta-analysis of few studies involving rare events. An empirical investigation from the CDSR confirmed the proposed WIP, that is $\mathcal{N}(0, 2.82^2)$. In a simulation study, the proposed method displayed desirable performance in terms of the investigated measures in comparison to fitting BNHM with a vague prior or using MLE. With the help of the `MetaStan`, the proposed method is easy to implement in R. In conclusion, I suggest the use of a WIP to deal with data sparsity in meta-analysis of few studies involving rare events.

3 Discussion

The idea of borrowing relevant information across different strata has a long history (Efron and Morris, 1975). Bayesian methods have been recommended for borrowing information in different phases of drug development (Davis and Leffingwell, 1990; Dixon and Simon, 1992) and in meta-analysis (Smith, Spiegelhalter, and A. Thomas, 1995; Sutton and Abrams, 2001). Besides trials themselves, different schedules in phase I or phase II trials can be treated as different strata. Borrowing information from different schedules in phase I or II trials is of great importance, since the sample sizes are small to moderate size and the number of strata is small in these trials. In this dissertation, I emphasize Bayesian methodology to analyze phase I and II trials with multiple schedules and the meta-analysis of few studies involving rare events. Here, I will outline the limitations of the proposed methods and provide some guidance for future work.

A novel Bayesian time-to-event model (TITE-PK) model for phase I dose-escalation trials with multiple schedules was introduced. One definition of the treatment schedule is the number of treatment cycles in the trial, which was used by Braun et al. (2007). I used an alternative definition of treatment schedules, that is the frequency of administration as suggested by Guo, Li, and Yuan (2016). The latter seems more relevant, since usually doctors will continue to administer a drug to the patients as long as patients benefit from the drug (Guo, Li, and Yuan, 2016). TITE-PK can also be used for phase I dose-escalation trials with multiple schedules using this definition of a schedule. To achieve this, the time point t^* must vary based on the number of treatment cycles, while assuming the frequency of administration f is shared between schedules. An important limitation of TITE-PK is the assumption of monotonicity of the exposure and the end-of-cycle 1 DLT probabilities. The monotonicity assumption implies completely ordered schedules, meaning that end-of-cycle 1 DLT probabilities increase with more frequent administrations given the same cumulative dose. The comparator POCRM relaxes the assumption of completely ordered schedules. However, TITE-PK yields robust performance against the violation of the completely ordered schedules by producing better or similar results in terms of the investigated metrics. For the sequential investigation of multiple schedules, we expect that the monotonicity assumption holds between schedules of the completed and the ongoing trials. In contrast, the comparator B-CRM approach explicitly models the heterogeneity between the completed and the ongoing trials. In the simulations, TITE-PK underperforms compared to B-CRM, when the scenario includes considerable heterogeneity between the completed and the ongoing trials in dose and DLT probability curves. Therefore, when there is a clear conflict of exposure and DLT probability relationships between the completed and the ongoing trials, B-CRM is a better choice than TITE-PK. One might consider an extension of TITE-PK

by including a meta-analytic-predictive (MAP) component to directly model heterogeneity between the completed and the ongoing trials. However, such extensions should be carefully investigated, since there might be identification problems due to data sparsity present in phase I trials. Another limitation of TITE-PK is that in the current implementation, only phase I trials with schedules of equal frequency of administrations can be analyzed. The model itself allows very different types of schedules, for instance a schedule with one day on, two days off etc. However, the **Stan** implementation of TITE-PK must be adapted for such schedules. Another important point is that TITE-PK is currently only suitable for a single drug, whereas combinations of drugs are becoming more and more popular. To extend TITE-PK for this purpose, possible drug interactions may need to be modeled.

I considered the estimation of dose-response models in phase II trials with multiple schedules. For this purpose, I propose a Bayesian hierarchical model in which certain parameters are treated as schedule specific random-effects, while others are assumed to be shared between schedules. Thus, the proposed model allows borrowing from different schedules instead of a static borrowing, that is using fixed weights to obtain schedule specific dose-response curves. The dynamic borrowing improves the accuracy of point estimates, while providing some robustness against the considerable heterogeneity in certain parameters between schedules. One disadvantage of the proposed model might be the parametrization of the model, since the overall mean parameter $\mu_{\mathrm{ED}_{50}}$ in Equation (2.6) does not have any meaningful interpretation. This is because the mean of schedules is not properly defined. To solve this, an alternative parametrization based on an asymmetric treatment of the schedule specific parameters can be used (Röver and Friede, 2020). Another disadvantage of the proposed model is that for some applications, there might be too much shrinkage, when it is not warranted. This may happen, when there exists an extreme schedule, thus the borrowing information for this schedule is not desirable. To overcome this, Neuenschwander, Wandel, et al. (2016) suggested the exchangeability-nonexchangeability (EXNEX) models, which avoid too much shrinkage for extreme strata (such as treatment schedule) in the dataset. EXNEX assumes that each strata is either exchangeable with some strata or nonexchangeable with any other strata, that is, an outlier. However, an EXNEX approach in our context must be tested well due to potential identification and convergence problems.

I proposed the use of weakly informative priors (WIP) for the treatment effect parameter in a binomial-normal hierarchical model (BNHM) for the meta-analysis of few studies involving rare events. The construction of a WIP for the treatment effect parameter was shown by assuming a normal prior with mean zero and assuming a priori interval for possible values. The constructed WIP is consistent with a re-analysis of Cochrane Database of Systematic Reviews (CDSR). In simulation studies, the proposed model has shown to have better performance in terms of accuracy of the point estimate and coverage of the interval estimate compared to alternatives. A main limitation of the simulation study is that the underlying data-generating process is also BNHM. Hence, the proposed model was not tested against the model misspecification. One might consider alternative parameterizations of the BNHM (Jackson et al., 2018) or a Poisson-normal hierarchical model as the data model. However, the main idea is still applicable for such models in which

relative treatment effects are modeled. A possible extension of the proposed method is network meta-analysis (NMA) involving rare events (Efthimiou et al., 2019). NMA is a generalization of the standard meta-analysis, which enables us to evaluate multiple treatments, even though they are not compared directly in a trial. The use of WIP for the treatment effect parameter in NMA can be especially useful, since in many NMA datasets, some treatment effect parameters are informed by only few trials. Although overall the number of trials can be considerably large, it does not guarantee that each treatment effect is informed by an adequate number of trials. This situation is similar to the standard meta-analysis of few trials. For example, for the BNHM model of NMA which was described in Günhan, Friede, and Held (2018), WIP of the treatment effect parameters for each treatment in the network can be constructed.

The use of **Stan** facilitates the implementation of the complicated Bayesian models which were proposed in the dissertation through MCMC methods. However, in order to use the developed R packages (`ModStan` and `MetaStan`), the user needs some knowledge about the convergence diagnostics of MCMC and **Stan**. Many practitioners are more familiar with the frequentist methods compared to MCMC methods. The novelty of MCMC diagnostics can be a barrier for some people to apply the proposed methods. For a gentle introduction to **Stan**, I refer to Sorensen, Hohenstein, and Vasishth (2016).

Bibliography

Albert, A and Anderson, JA (1984). "On the existence of maximum likelihood estimates in logistic regression models". In: *Biometrika* 71.1, pp. 1–10. DOI: 10.2307/2336390.

Amgen (2019). *Efficacy and safety of AMG 570 in subjects with active systemic lupus erythematosus*. Identification No. NCT04058028. Updated May, 2020. Accessed September, 2020. URL: https://clinicaltrials.gov/ct2/show/NCT04058028.

Angevin, E et al. (2013). "Phase I study of dovitinib (TKI258), an oral FGFR, VEGFR, and PDGFR inhibitor, in advanced or metastatic renal cell carcinoma". In: *Clinical Cancer Research* 19.5, pp. 1257–1268. DOI: 10.1158/1078-0432.CCR-12-2885.

Babb, J, Rogatko, A, and Zacks, S (1998). "Cancer phase I clinical trials: Efficient dose escalation with overdose control". In: *Statistics in Medicine* 17.10, pp. 1103–1120. DOI: 10.1002/(SICI)1097-0258(19980530)17:10<1103::AID-SIM793>3.0.CO;2-9.

Baeten, D et al. (2013). "Anti-interleukin-17A monoclonal antibody secukinumab in treatment of ankylosing spondylitis: a randomised, double-blind, placebo-controlled trial". In: *The Lancet* 382.9906, pp. 1705–1713. DOI: https://doi.org/10.1016/S0140-6736(13)61134-4.

Bates, D et al. (2015). "Fitting linear mixed-effects models using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Besse, B et al. (2014). "A phase Ib dose-escalation study of everolimus combined with cisplatin and etoposide as first-line therapy in patients with extensive-stage small-cell lung cancer". In: *Annals of Oncology* 25.2, pp. 505–511. DOI: 10.1093/annonc/mdt535.

Böhning, D, Mylona, K, and Kimber, A (2015). "Meta-analysis of clinical trials with rare events." In: *Biometrical Journal* 57.4, pp. 633–648. DOI: 10.1002/bimj.201400184.

Bornkamp, B (2012). "Functional uniform priors for nonlinear modeling". In: *Biometrics* 68.3, pp. 893–901. DOI: 10.1lll/j.1541-0420.2012.017.

— (2014). "Practical considerations for using functional uniform prior distributions for dose-response estimation in clinical trials". In: *Biometrical Journal* 56.6, pp. 947–962. DOI: 10.1002/bimj.201300138.

Bradburn, MJ et al. (2007). "Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events". In: *Statistics in Medicine* 26.1, pp. 53–77. DOI: 10.1002/sim.2528.

Braun, TM et al. (2007). "Simultaneously optimizing dose and schedule of a new cytotoxic agent". In: *Clinical Trials* 4.2, pp. 113–124. DOI: 10.1177/1740774507076934.

Carpenter, B et al. (2017). "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 76.1, pp. 1–32. DOI: 10.18637/jss.v076.i01.

Celgene (2018). *Vidaza (Azacitidine): Highlights of prescribing information*. Updated July, 2018. Accessed February, 2020. Summit, New Jersey, USA: Author. URL: https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/050794s031lbl.pdf.

Chuang-Stein, C and Kirby, S (2017). *Quantitative Decisions in Drug Development*. Cham, Switzerland: Springer International Publishing. DOI: 10.1007/978-3-319-46076-5.

Cochrane (2020). *Cochrane Database of Systematic Reviews*. URL: https://www.cochranelibrary.com.

Crins, ND et al. (2014). "Interleukin-2 receptor antagonists for pediatric liver transplant recipients: A systematic review and meta-analysis of controlled studies". In: *Pediatric Transplantation* 18.8, pp. 839–850. DOI: 10.1111/petr.12362.

Cunanan, KM and Koopmeiners, JS (2017). "A Bayesian adaptive phase I-II trial design for optimizing the schedule of therapeutic cancer vaccines". In: *Stat Med* 36.1, pp. 43–53. DOI: 10.1002/sim.7087.

Davis, CE and Leffingwell, DP (1990). "Empirical Bayes estimates of subgroup effects in clinical trials". In: *Controlled Clinical Trials* 11.1, pp. 37–42. DOI: https://doi.org/10.1016/0197-2456(90)90030-6.

de Lima, M et al. (2010). "Maintenance therapy with low-dose azacitidine after allogeneic hematopoietic stem cell transplantation for recurrent acute myelogenous leukemia or myelodysplastic syndrome". In: *Cancer* 116.23, pp. 5420–5431. DOI: 10.1002/cncr.25500.

Demetri, GD et al. (2009). "A phase I study of single-agent nilotinib or in combination with imatinib in patients with imatinib-resistant gastrointestinal stromal tumors". In: *Clinical Cancer Research* 15.18, pp. 5910–5916. DOI: 10.1158/1078-0432.CCR-09-0542.

DerSimonian, R and Laird, N (1986). "Meta-analysis in clinical trials". In: *Controlled Clinical Trials* 7.3, pp. 177–188. DOI: https://doi.org/10.1016/0197-2456(86)90046-2.

Dixon, DO and Simon, R (1992). "Bayesian subset analysis in a colorectal cancer clinical trial". In: *Statistics in Medicine* 11.1, pp. 13–22. DOI: 10.1002/sim.4780110104.

Efron, B and Morris, C (1975). "Data analysis using Stein's estimator and its generalizations." In: *Journal of the American Statistical Association* 70.350, pp. 311–319. URL: http://www.jstor.org/stable/2285814.

Efthimiou, O et al. (2019). "Network meta-analysis of rare events using the Mantel-Haenszel method". In: *Statistics in Medicine* 38.16, pp. 2992–3012. DOI: 10.1002/sim.8158.

Esaki, T et al. (2019). "Phase I dose-escalation study of capmatinib (INC280) in Japanese patients with advanced solid tumors". In: *Cancer Science* 110.4, pp. 1340–1351. DOI: 10.1111/cas.13956.

European Medicines Agency (2006). *Guideline on clinical trials in small populations.* Updated July, 2006. Accessed August, 2020. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-trials-small-populations_en.pdf.

Feller, C et al. (2017). "Optimal designs for dose response curves with common parameters". In: *Annals of Statistics* 45.5, pp. 2102–2132. DOI: 10.1214/16-AOS1520.

Firth, D (1993). "Bias reduction of maximum likelihood estimates". In: *Biometrika* 80.1, pp. 27–38. DOI: 10.2307/2336755.

Friede, T et al. (2017). "Meta-analysis of few small studies in orphan diseases". In: *Research Synthesis Methods* 8.1, pp. 79–91. DOI: 10.1002/jrsm.1217.

Galapagos NV (2020). *A study to assess efficacy, safety, tolerability and pharmacokinetics (PK)/pharmacodynamics (PD) of MOR106 in subjects with moderate to severe atopic dermatitis (IGUANA)*. Identification No. NCT03568071. Updated March, 2020. Accessed September, 2020. URL: https://clinicaltrials.gov/ct2/show/NCT03568071?cond=NCT03568071&draw=2&rank=1.

Gelman, A (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)". In: *Bayesian Analysis* 1.3, pp. 515–534. DOI: 10.1214/06-BA117A.

Gelman, A, Carlin, JB, et al. (2013). *Bayesian Data Analysis*. 3rd ed. Boca Raton, Florida: CRC Press. DOI: https://doi.org/10.1201/b16018.

Gelman, A, Jakulin, A, et al. (2008). "A weakly informative default prior distribution for logistic and other regression models". In: *The Annals of Applied Statistics* 2.4, pp. 1360–1383. DOI: 10.1214/08-AOAS191.

Greenland, S and Mansournia, MA (2015). "Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions". In: *Statistics in Medicine* 34.23, pp. 3133–3143. DOI: 10.1002/sim.6537.

Günhan, BK, Friede, T, and Held, L (2018). "A design-by-treatment interaction model for network meta-analysis and meta-regression with integrated nested Laplace approximations." In: *Research Synthesis Methods* 9.2, pp. 179–194. DOI: 10.1002/jrsm.1285.

Günhan, BK, Meyvisch, P, and Friede, T (2020). "Shrinkage estimation for dose-response modeling in phase II trials with multiple schedules". In: *Statistics in Biopharmaceutical Research*, pp. 1–15. DOI: https://doi.org/10.1080/19466315.2020.1850519.

Günhan, BK, Röver, C, and Friede, T (2020). "Random-effects meta-analysis of few studies involving rare events". In: *Research Synthesis Methods* 11.1, pp. 74–90. DOI: 10.1002/jrsm.1370.

Günhan, BK, Weber, S, and Friede, T (2020). "A Bayesian time-to-event pharmacokinetic model for phase I dose-escalation trials with multiple schedules". In: *Statistics in Medicine* 39.27, pp. 3986–4000. DOI: 10.1002/sim.8703.

Günhan, BK, Weber, S, Seroutou, A, et al. (2020). *A Bayesian time-to-event pharmacokinetic model for sequential phase I dose-escalation trials with multiple schedules*. Updated August, 2020. Accessed August, 2020. URL: https://arxiv.org/abs/1811.09433.

Guo, B, Li, Y, and Yuan, Y (2016). "A dose-schedule finding design for phase I-II clinical trials". In: *Journal of Royal Statistics Society: Series C* 65.2, pp. 259–272. DOI: 10.1111/rssc.12113.

Higgins, JPT and Green, S (eds) (2008). "Cochrane handbook for systematic reviews of interventions". In: Chichester: UK: Wiley. DOI: 10.1002/9780470712184.

Higgins, JPT, Thompson, SG, and Spiegelhalter, DJ (2009). "A re-evaluation of random-effects meta-analysis." In: *Journal of Royal Statistical Society Series A (Statistics in Society)* 172.1, pp. 137–159. DOI: 10.1111/j.1467-985X.2008.00552.x.

Jackson, D et al. (2018). "A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio". In: *Statistics in Medicine* 37.7, pp. 1059–1085. DOI: 10.1002/sim.7588.

Jones, HE et al. (2011). "Bayesian models for subgroup analysis in clinical trials". In: *Clinical Trials* 8.2, pp. 129–143.

Kalbfleisch, JD and Prentice, RL (2002). *The Statistical Analysis of Failure Time Data*. New York, NY: John Wiley & Sons. DOI: 10.1002/9781118032985.

Kallen, A (2007). *Computational Pharmacokinetics*. Boca Raton, Florida: CRC Press. DOI: 10.1201/9781420060669.

Kuss, O (2015). "Statistical methods for meta-analyses including information from studies without any events–add nothing to nothing and succeed nevertheless". In: *Statistics in Medicine* 34.7, pp. 1097–1116. DOI: 10.1002/sim.6383.

Le Tourneau, C, Lee, JJ, and Siu, LL (2009). "Dose escalation methods in phase I cancer clinical trials". In: *Journal of the National Cancer Institute* 101.10, pp. 708–720. DOI: 10.1093/jnci/djp079.

Liu, S et al. (2015). "Bridging continual reassessment method for phase I clinical trials in different ethnic populations". In: *Statistics in Medicine* 34.10, pp. 1681–1694. DOI: 10.1002/sim.6442.

Mantel, N and Haenszel, W (1959). "Statistical aspects of the analysis of data from retrospective studies of disease". In: *Journal of the National Cancer Institute* 22.4, pp. 719–748. DOI: 10.1093/jnci/22.4.719.

Möllenhoff, K, Bretz, F, and Dette, H (2020). "Equivalence of regression curves sharing common parameters." In: *Biometrics* 76.2, pp. 518–529. DOI: 10.1111/biom.13149.

Murphy, P et al. (2017). "Lemborexant, a dual orexin receptor antagonist (DORA) for the treatment of insomnia disorder: Results from a Bayesian, adaptive, randomized, double-blind, placebo-controlled study". In: *Journal of Clinical Sleep Medicine* 13.11, pp. 1289–1299. DOI: 10.5664/jcsm.6800.

Naing, A et al. (2020). "A first-in-human phase 1 dose escalation study of spartalizumab (PDR001), an anti–PD-1 antibody, in patients with advanced solid tumors". In: *Journal for ImmunoTherapy of Cancer* 8.1. DOI: 10.1136/jitc-2020-000530.

Neuenschwander, B, Matano, A, et al. (2015). "A Bayesian industry approach to phase I combination trials in oncology". In: *Statistical Methods in Drug Combination Studies*. Boca Raton, Florida: CRC Press. DOI: 10.1201/b17965.

Neuenschwander, B, Wandel, S, et al. (2016). "Robust exchangeability designs for early phase clinical trials with multiple strata". In: *Pharmaceutical Statistics* 15.2, pp. 123–134. DOI: 10.1002/pst.1730.

Novartis (2012). *Afinitor (Everolimus): Highlights of prescribing information*. Updated July, 2012. Accessed April, 2020. Basel, Switzerland: Author. URL: https://www.accessdata.fda.gov/drugsatfda_docs/label/2012/022334s016lbl.pdf.

O'Quigley, J, Pepe, M, and Fisher, L (1990). "Continual reassessment method: A practical design for phase 1 clinical trials in cancer." In: *Biometrics* 46.1, pp. 33–48. URL: https://doi.org/10.2307/2531628.

Ollier, A et al. (2020). "An adaptive power prior for sequential clinical trials – Application to bridging studies". In: *Statistical Methods in Medical Research* 29.8, pp. 2282–2294. DOI: 10.1177/0962280219886609.

Pfizer (2017). *Monthly And twice monthly subcutaneous dosing of PF-04950615 (RN316) In hypercholesterolemic subjects on a statin*. Identification No. NCT01592240. Updated December, 2017. Accessed August, 2020. URL: https://clinicaltrials.gov/ct2/show/results/NCT01592240?cond=NCT01592240&draw=2&rank=1&view=results.

Price, K and LaVange, L (2014). "Bayesian methods in medical product development and regulatory reviews". In: *Pharmaceutical Statistics* 13.1, pp. 1–2. DOI: 10.1002/pst.1608.

Röver, C and Friede, T (2020). "Dynamically borrowing strength from another study through shrinkage estimation". In: *Statistical Methods in Medical Research* 29.1, pp. 293–308. DOI: 10.1177/0962280219833079.

Ruberg, SJ (1995). "Dose response studies I. some design considerations". In: *Journal of Biopharmaceutical Statistics* 5.1, pp. 1–14. DOI: 10.1080/10543409508835096.

Shah, NP et al. (2008). "Intermittent target inhibition with dasatinib 100 mg once daily preserves efficacy and improves tolerability in imatinib-resistant and -intolerant chronic-phase chronic myeloid leukemia". In: *Journal of Clinical Oncology* 26.19, pp. 3204–3212. DOI: 10.1200/JCO.2007.14.9260.

Smith, TC, Spiegelhalter, DJ, and Thomas, A (1995). "Bayesian approaches to random-effects meta-analysis: A comparative study". In: *Statistics in Medicine* 14.24, pp. 2685–2699. DOI: 10.1002/sim.4780142408.

Sorensen, T, Hohenstein, S, and Vasishth, S (2016). "Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists". In: *The Quantitative Methods for Psychology* 12.3, pp. 175–200. DOI: 10.20982/tqmp.12.3.p175.

Springate, D (2018). *Cochrane_scraper*. version 1.1.0. Updated July, 2014. Accessed April, 2018. URL: https://doi.org/10.5281/zenodo.10782.

Sutton, AJ and Abrams, KR (2001). "Bayesian methods in meta-analysis and evidence synthesis". In: *Statistical Methods in Medical Research* 10.4, pp. 270–303. DOI: doi:10.1177/096228020101000404.

Thaçi, D et al. (2016). "Efficacy and safety of dupilumab in adults with moderate-to-severe atopic dermatitis inadequately controlled by topical treatments: a randomised, placebo-controlled, dose-ranging phase 2b trial". In: *Lancet* 387.10013, pp. 40–52. DOI: 10.1016/S0140-6736(15)00388-8.

Thall, PF et al. (2013). "Using joint utilities of the times to response and toxicity to adaptively optimize schedule-dose regimes". In: *Biometrics* 69.3, pp. 673–682. DOI: 10.1111/biom.12065.

Thomas, N, Sweeney, K, and Somayaji, V (2014). "Meta-analysis of clinical dose-response in a large drug development portfolio." In: *Statistics in Biopharmaceutical Research* 6.4, pp. 302–317. DOI: 10.1080/19466315.2014.924876.

Turner, RM, Davey, J, et al. (2012). "Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews". In: *International Journal of Epidemiology* 41.3, pp. 818–827. DOI: 10.1093/ije/dys041.

Turner, RM, Jackson, D, et al. (2015). "Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis." In: *Statistics in Medicine* 34.6, pp. 984–998. DOI: 10.1002/sim.6381.

U.S. Food and Drug Administration (2010). *Guidance for the use of Bayesian statistics in medical device clinical trials.* Updated February, 2010. Accessed September, 2020. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials.

— (2014). *Complex issues in developing drugs and biological products for rare diseases and accelerating the development of therapies for pediatric rare diseases including strategic plan: Accelerating the development of therapies for pediatric rare diseases.* Updated July, 2014. Accessed April, 2020. URL: https://www.fda.gov/media/89051/download.

— (2018). *Meta-analyses of randomized controlled clinical trials to evaluate the safety of human drugs or biological products.* Updated November, 2018. Accessed September, 2020. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/meta-analyses-randomized-controlled-clinical-trials-evaluate-safety-human-drugs-or-biological.

— (2019). *Adaptive design clinical trials for drugs and biologics guidance for industry.* Updated December, 2019. Accessed September, 2020. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.

Vandermeer, B et al. (2009). "Meta-analyses of safety data: a comparison of exact versus asymptotic methods". In: *Statistical Methods in Medical Research* 18.4, pp. 421–432. DOI: 10.1177/0962280208092559.

Vehtari, A, Gelman, A, and Gabry, J (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and Computing* 27.5, pp. 1413–1432. DOI: 10.1007/s11222-016-9696-4.

Viele, K et al. (2014). "Use of historical control data for assessing treatment effects in clinical trials". In: *Pharmaceutical Statistics* 13.1, pp. 41–54. DOI: 10.1002/pst.1589.

Wages, NA (2017). "Dose–schedule finding in early-phase clinical trials". In: *Handbook of Methods for Designing, Monitoring, and Analyzing Dose-Finding Trials*. Boca Raton, Florida: CRC Press. DOI: https://doi.org/10.1201/9781315151984.

Wages, NA, O'Quigley, J, and Conaway, MR (2014). "Phase I design for completely or partially ordered treatment schedules". In: *Statistics in Medicine* 33.4, pp. 569–579. DOI: 10.1002/sim.5998.

Wagner, LM et al. (2010). "Phase I trial of two schedules of vincristine, oral irinotecan, and temozolomide (VOIT) for children with relapsed or refractory solid tumors: A Children's Oncology Group phase I consortium study". In: *Pediatric Blood & Cancer* 54.4, pp. 538–545. DOI: 10.1002/pbc.22407.

A Original articles

A.1 A Bayesian time-to-event pharmacokinetic model for phase I dose-escalation trials with multiple schedules

Published version is publicly available from https://doi.org/10.1002/sim.8703.

Received: 23 November 2018 | Revised: 21 April 2020 | Accepted: 30 June 2020

DOI: 10.1002/sim.8703

RESEARCH ARTICLE

Statistics in Medicine WILEY

A Bayesian time-to-event pharmacokinetic model for phase I dose-escalation trials with multiple schedules

Burak Kürsad Günhan[1] | **Sebastian Weber**[2] | **Tim Friede**[1]

[1]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[2]Advanced Exploratory Analytics, Novartis Pharma AG, Basel, Switzerland

Correspondence
Burak Kürsad Günhan, Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany.
Email: burak.gunhan@med.uni-goettingen.de

Phase I dose-escalation trials must be guided by a safety model in order to avoid exposing patients to unacceptably high risk of toxicities. Traditionally, these trials are based on one type of schedule. In more recent practice, however, there is often a need to consider more than one schedule, which means that in addition to the dose itself, the schedule needs to be varied in the trial. Hence, the aim is finding an acceptable dose-schedule combination. However, most established methods for dose-escalation trials are designed to escalate the dose only and ad hoc choices must be made to adapt these to the more complicated setting of finding an acceptable dose-schedule combination. In this article, we introduce a Bayesian time-to-event model which takes explicitly the dose amount and schedule into account through the use of pharmacokinetic principles. The model uses a time-varying exposure measure to account for the risk of a dose-limiting toxicity over time. The dose-schedule decisions are informed by an escalation with overdose control criterion. The model is formulated using interpretable parameters which facilitates the specification of priors. In a simulation study, we compared the proposed method with an existing method. The simulation study demonstrates that the proposed method yields similar or better results compared with an existing method in terms of recommending acceptable dose-schedule combinations, yet reduces the number of patients enrolled in most of scenarios. The `R` and `Stan` code to implement the proposed method is publicly available from Github (https://github.com/gunhanb/TITEPK_code).

KEYWORDS

multiple schedules, pharmacokinetic models, phase I dose-escalation trials, Stan

1 | INTRODUCTION

In a phase I trial, a treatment plan includes the amount of drug to be given a patient, known as the dose, and the times when it is given, known as the schedule. Phase I dose-escalation trials traditionally include only one schedule while varying the dose among patients. However, in medical practice, there is often a need to consider different schedules, for example, a dose given once a day or once a week, within a phase I trial. Many established methods[1,2] are only designed for varying the drug amount since time is not taken into account in the models. These approaches require ad hoc adjustments

like scaling the dose to accommodate these more complex designs. However, varying treatment schedules necessitates to take time into account in the model as the pharmacokinetic (PK) properties of a drug become relevant whenever the treatment schedule is varied.

The aim of a phase I dose-escalation trial with multiple schedules is finding an acceptable dose and schedule combination. The dose-escalation must be guided by a safety model in order to avoid exposing patients to unacceptably high risk of toxicities. What defines an acceptable dose-schedule combination depends on the drug development strategy. In oncology, the efficacy is sought to be maximized at the cost of tolerating safety events, referred to as dose-limiting toxicities (DLTs). Therefore, one seeks in oncology a so-called maximum tolerated dose-schedule combination (MTC) at which an acceptable rate of DLT events is expected to occur.

Different definitions of treatment schedule have been used in the literature.[3,4] When the aim is to optimize the number of cycles to treat patients, the schedule is defined as the number of treatment cycles. An alternative definition is the frequency (timing) of administration within a cycle with a given total dose per cycle. Guo et al[3] argued that this definition of schedule seems more relevant in practical terms, since physicians will usually continue to treat patients as long as patients appear to benefit from the treatment. A more interesting point for the physicians is how frequently the treatment should be administered. We agree with this reasoning and hence adopted this second definition of schedule in the article.

There are different methods suggested for determining the MTC. Using the definition of a schedule as the number of treatment cycles, Braun et al[5] developed a time-to-event model which simultaneously optimizes the dose and the schedule. Zhang and Braun[6] extended this method to incorporate adaptive variations to dose-schedule assignments within patients as the trial proceeds. Wages et al[7] introduce a dose-schedule finding design, the partial order continual reassessment method (POCRM), which relaxes the assumption of completely ordered schedules, that is, for a given dose, DLT probabilities are completely ordered in terms of different schedules. Wages et al[7] used the second definition of the schedule.[4] Furthermore, Li et al,[8] Thall et al,[9] Guo et al,[3] and Cunanan and Koopmeiners[10] suggested dose-schedule finding methods that jointly models efficacy and toxicity in the context of dose-schedule combination designs.

We introduce an alternative model, a *time-to-event pharmacokinetic (TITE-PK) model* henceforth referred as TITE-PK, that uses PK principles to introduce an exposure measure. Consequently, TITE-PK is an exposure-response model[11] which usually uses more information than a standard dose-response model, such as kinetic drug properties. Formally, a pseudo-PK model is used to define a time-varying exposure measure, which constitutes a time-varying Poisson process describing the DLT event process. TITE-PK utilizes data on the exact treatment schedule and time-to-first DLT in a fully Bayesian model-based approach following the spirit of Cox et al[12] To inform dose-schedule decisions, TITE-PK uses an adapted escalation with overdose control (EWOC)[13] criterion. This requires that for a given dose-schedule combination, the probability for a DLT occurred within the first cycle must not exceed the maximal admissible DLT probability by a prespecified feasibility bound. In the proposed model, PK analysis and safety analysis are not combined as is done, for example, by Ursino et al[14] Instead we use a pseudo-PK model in TITE-PK which can be seen as a kinetic-pharmacodynamic model, see, for example, Jacqmin et al[15] and Jacons et al[16]

We provide simulations comparing the performance of our proposed model to the POCRM method, which is in the spirit of the continual reassessment design (CRM).[1] POCRM was originally developed for drug combination trials,[17] and later extended to phase I trials with multiple schedules.[7] The simulation study is motivated by the Vidaza trial.[18] Vidaza is a cytotoxic drug that is used for the treatment of a blood cell disease, known as myelodysplastic syndrome, that often develops into acute myelogeneos leukemia. The Vidaza trial (ClinicalTrials.gov identifier: NCT01080664) investigated four different schedules and three doses, and thus, is an example of a dose-schedule finding problem.[5] The `R` and `Stan` code for the implementation of the proposed TITE-PK model is available from Github (https://github.com/gunhanb/TITEPK_code).

This article is structured as follows. In Section 2, we introduce the proposed TITE-PK model. In Section 3, the performance of TITE-PK and POCRM are compared in a simulation study. We close with a discussion and a conclusion.

2 | THE PROPOSED MODEL: TITE-PK

The time-to-first DLT is modeled using a time-varying (nonhomogeneous) Poisson process. A time-varying Poisson process can be defined using the instantaneous hazard function ($h(t)$) for a DLT occurring at time t. The hazard function corresponds to the probability that a patient experiences a DLT in the time interval $(t, t + \delta t]$ given that they did not

experience a DLT until time t. The hazard is modeled as a time-dependent function directly proportional to an exposure measure of the drug ($E(t)$) as Reference 12

$$h(t) = \beta\ E(t), \tag{1}$$

where β is the proportionality parameter to estimate. Here, the exposure measure refers to the drug concentration as in an exposure-response model,[11] and the calculation of $E(t)$ will be explained in Section 2.1. Furthermore, if we integrate both sides of Equation (1) from time 0 up to time t, we obtain

$$H(t) = \beta\ \mathrm{AUC}_E(t), \tag{2}$$

where $\mathrm{AUC}_E(t)$ is the area under the curve of the exposure measure over time and $H(t)$ is the cumulative hazard function, respectively. From event history analysis,[19] we know that the probability density for an event to occur at time-point t is

$$f(t) = h(t)\ \exp(-H(t)) \tag{3}$$

and the survivor function for the event to occur past some time-point t is given by

$$S(t) = P(T > t) = \exp(-H(t)), \tag{4}$$

where T denotes the event time. In the following, we use C_j to denote the censoring time of patient j. Accordingly, TITE-PK is able to account for the partial information from subjects still in the follow-up (censored patients) like TITE-CRM.[20] By contrast to TITE-CRM, we restrict the follow-up period for all patients to cycle 1 only, which is a conventional approach. Thus, all patients without a DLT up to the end of cycle 1 will be censored at the end of cycle 1, $C_j = t^*$. Furthermore, we denote with δ_j an event indicator which is set to 0 for censored events and 1 for DLT events. The overall likelihood can be written as

$$L(T, C|\beta) = \prod_{j=1}^{J} f(T_j|\beta)^{\delta_j}\ S(C_j|\beta)^{(1-\delta_j)}, \tag{5}$$

where J is the total number of the patients. Now, we discuss the exposure measure of the drug.

2.1 | Pseudo-PK model

The proposed exposure model in the TITE-PK model does not rely on measured drug concentration data, as this data is not routinely available in a form that it may be used directly in the model to support dose-schedule decisions in a timely manner. For this reason, PK is considered as latent variable which we refer to *pseudo*-PK. The pseudo-PK is used to account for the dosing history and the expected accumulation in exposure over time that ultimately drive pharmacological responses, including safety. The main purpose of the proposed pseudo-PK model is to account for the natural "waxing and waning" of exposure observed after dosing of drug. This pseudo-PK model has a *central* compartment into which the drug is administered and accounts for drug elimination as a linear first-order process; that is, the elimination rate is proportional to the amount of drug in the compartment[21]

$$\frac{dC(t)}{dt} = -k_e\ C(t), \tag{6}$$

where $C(t)$ is the concentration of drug in the central compartment and k_e is the elimination rate constant, which is given by log(2) divided by the elimination half-life. As the volume of the central compartment cannot be identified for a latent *pseudo*-PK, we set it by convention to unity.

To account for delays between the instantaneous drug concentration in the central compartment and the concentration during the pharmacodynamic effect, we use a so-called effect compartment[21]

$$\frac{dC_{\mathrm{eff}}(t)}{dt} = k_{\mathrm{eff}}\ (C(t) - C_{\mathrm{eff}}(t)). \tag{7}$$

Here, $C_{\text{eff}}(t)$ is the drug concentration in the effect compartment and k_{eff} is the PK parameter which governs the delay between the concentration in the central compartment ($C(t)$) and the concentration in the effect compartment ($C_{\text{eff}}(t)$). Note that the solutions of Equations (6) and (7) are the same up to reparametrization for a one compartment model with first-order absorption, which would be one way to model oral absorption. The parameters k_e and k_{eff} are assumed to be known from previous PK analyses, for example, from preclinical experiments. The model is conditioned on the PK parameters from previous analyses, thus uncertainty of the PK parameters in the estimation is ignored. A procedure to calculate an estimate for k_{eff} using the cycle duration and the absorption rate is described in Section 3.1 for the Vidaza trial.

The ordinary differential equations (ODE) (6) and (7) account for dosing over time through administration into the central compartment. The analytical solution to the ODE system for multiple doses is obtained through the use of the superposition principle which holds for linear ODE systems (see, eg, Reference 22). This model in principle can account for the history of any treatment schedule over time. In order to simplify the notation, we restrict ourselves here to regular treatment schedules which have a dosing frequency f (in units of 1/h), start at time $t = 0$h and use the same dose amount d for all dosing events. With these simplifications (in notation) the solution to the above ODE system is

$$C_{\text{eff}}(t|d,f) = d \sum_{i=0}^{\infty} \Theta\left(t - \frac{i}{f}\right) \frac{k_{\text{eff}}}{k_{\text{eff}} - k_e} \left(e^{-k_e\left(t-\frac{i}{f}\right)} - e^{-k_{\text{eff}}\left(t-\frac{i}{f}\right)}\right), \tag{8}$$

where Θ denotes the Heaviside step function (or unit step function).

To facilitate meaningful interpretation of the parameter β and hence to help prior specification, the exposure measure $E(t)$ is obtained by scaling $C_{\text{eff}}(t)$ using a reference dose-schedule combination including a reference dose (d^*) and a reference dosing frequency (f^*) at the end of cycle 1 (t^*) such that

$$E(t|d,f) = \frac{C_{\text{eff}}(t|d,f)}{\int_0^{t^*} C_{\text{eff}}(t|d^*,f^*)\, dt}$$

$$\text{AUC}_E(t^*|d^*,f^*) = \int_0^{t^*} E(t|d^*,f^*)\, dt = 1.$$

This is analogous to using a reference dose in the Bayesian Logistic Regression Model[2] which is a two parameter version of the CRM design and uses the EWOC criterion for dose-escalation decisions.

2.2 | Informing dose-schedule decisions

To inform dose-schedule decisions, TITE-PK uses an adapted EWOC criterion. The probability that a patient experiences at least one DLT within the first cycle (shortly the end-of-cycle 1 DLT probability) given the dose-schedule combination with dose d and frequency f, $P(T \le t^*|d,f)$, is our measure of interest.

The end-of-cycle 1 DLT probabilities are classified into three categories as follows

(i) $P(T \le t^*|d,f) < 0.16$ Underdosing (UD)
(ii) $0.16 \le P(T \le t^*|d,f) \le 0.33$ Targeted toxicity (TT)
(iii) $P(T \le t^*|d,f) > 0.33$ Overdosing (OD)

Dose-schedule decisions are informed using the OD probability of the dose-schedule combination d and f. The EWOC criterion is fulfilled, if $P(P(T \le t^*|d,f) > 0.33)$ is smaller than the prespecified feasibility bound a. Among the dose-schedule combinations which fulfill the EWOC criterion, the combination which has the lowest $\text{AUC}_E(t^*)$ is recommended by TITE-PK. This is analogous to recommending the lowest dose amount in the "standard" phase I dose-escalation methods. When $\text{AUC}_E(t^*)$ of eligible combinations are exactly the same, one of the dose-schedule combinations can be chosen randomly. In this article, we use $a = 0.25$, which was suggested by Babb et al,[13] and also $a = 0.50$ to investigate the sensitivity of the results to the choice of a in our simulations. The higher (lower) value of the feasibility bound make it easier (harder) to escalate to the next dose and schedule combinations, resulting in more (less) aggressive dose-schedule decisions.

By using the relationship between $P(T \leq t^*|d,f) = 1 - P(T > t^*|d,f)$ and combining Equation (4) with (2) it follows that

$$\begin{aligned} P(T > t^*|d,f) &= \exp(-H(t^*|d,f)) = 1 - P(T \leq t^*|d,f) \\ &\Leftrightarrow \log(H(t^*|d,f)) = \log(-\log(1 - P(T \leq t^*|d,f))) = \text{cloglog}(P(T \leq t^*|d,f)), \end{aligned} \tag{9}$$

where $\text{cloglog}(x) = \log(-\log(1-x))$.

Since the cumulative hazard $H(t|d,f)$ is set proportional, see Equation (2), to the area under curve of the exposure metric $\text{AUC}_E(t|d,f)$ this leads to

$$\text{cloglog}(P(T \leq t^*|d,f)) = \log(\beta) + \log(\text{AUC}_E(t^*|d,f)). \tag{10}$$

For the reference dose-schedule combination with dose d^* and dosing frequency f^* the AUC of the exposure measure up to the reference time-point is unity, $\text{AUC}_E(t^*|d^*,f^*) = 1$, such that $\text{cloglog}(P(T \leq t^*|d^*,f^*)) = \log(\beta)$ holds. This highlights the importance of the reference dose-schedule combination to specify the prior for the parameter β.

As is apparent from Equation (2), TITE-PK assumes that the end-of-cycle 1 DLT probability is a monotonic function of the exposure measure. Moreover, the pseudo-PK model assumes a linear first-order process (linear PK), meaning that increases in drug exposure are linearly related to increases in administered doses. Consequently, $\text{AUC}_E(t^*|d,f)$ is directly proportional to d. Thus, the assumption of the monotonicity of the exposure and the end-of-cycle 1 DLT probability implies the assumption of the monotonicity of the dose and the end-of-cycle 1 DLT probability. However, we will see in the simulations that the performance of the model is robust to some extent to violations of the monotonicity assumption.

2.3 | Software implementation

The proposed model TITE-PK is implemented in `Stan`[23] via **rstan** `R` package. The corresponding code for the implementation of the TITE-PK model is available from Github (https://github.com/gunhanb/TITEPK_code). Four parallel chains of 1000 MCMC iterations after warm-up of 1000 iterations are generated. Convergence diagnostics are checked using the Gelman-Rubin statistics[24] and traceplots.

3 | SIMULATION STUDY

In order to assess the performance of the TITE-PK and to compare with the POCRM under different true dose-DLT profiles with multiple schedules, various scenarios are investigated in a simulation study.

3.1 | Simulation settings

The scenarios considered in the article were also investigated by Wages et al[7] These are motivated by the Vidaza example.[18] As in the Vidaza trial, the scenarios investigated four different schedules (A, B, C, D) and three doses (8, 16, 24 mg/m^2). In the Vidaza trial, different treatment schedules correspond to the number of cycles the drug is administered with a prespecified frequency of administration. More specifically, four schedules are 1, 2, 3, or 4 cycles, each with 5 days of drug administration and 25 days of rest. As explained in the introduction, here we use another definition of the schedule, the timing of the administration within the first cycle only. To mimic the nested schedules of the Vidaza trial, we chose the four schedules A, B, C, and D as dosing frequency of 192, 96, 48, and 24 hours, respectively. The cycle length is taken as 28 days ($t^* = 28$ days). The reference dose and the reference dosing frequency are determined using 24 mg ($d^* = 24$ mg) and Schedule B ($f^* = 1/96$ 1/h).

The scenarios were carefully chosen to reflect a range of clinically relevant scenarios. These are summarized in Table 1 and illustrated by Figure 1. Scenario 1 does not include any dose-schedule combination in the OD interval, whereas all combinations are in the OD interval in Scenario 2. Scenarios 3 and 4 are scenarios in which DLT probabilities are spread across UD, TT, and OD intervals. Five and three dose-schedule combinations are in the TT interval in Scenarios 3 and 4, respectively. In Scenario 5, there is only one dose-schedule combination in the TT interval. In addition, there is a heavy

TABLE 1 Toxicity scenarios for the dose-schedule combination in the simulation study

	Doses in mg/m^2					
Schedule	**8**	**16**	**24**	**8**	**16**	**24**
	Scenario 1			Scenario 2		
A	0.05	0.07	0.11	0.50	0.54	0.58
B	0.09	0.12	0.18	0.53	0.60	0.65
C	0.16	0.18	**0.23**	0.55	0.65	0.75
D	**0.22**	**0.26**	**0.30**	0.57	0.73	0.78
	Scenario 3			Scenario 4		
A	0.03	0.14	**0.28**	0.03	0.15	**0.30**
B	0.09	**0.21**	**0.40**	0.12	**0.30**	0.50
C	0.18	**0.32**	0.54	**0.30**	0.50	0.60
D	**0.31**	0.45	0.62	0.50	0.60	0.75
	Scenario 5			Scenario 6		
A	0.01	0.10	0.50	0.05	0.07	0.11
B	0.03	**0.30**	0.55	0.16	0.18	**0.23**
C	0.05	0.50	0.60	0.09	0.12	0.18
D	0.10	0.60	0.70	**0.22**	**0.26**	**0.30**
	Scenario 7					
A	0.10	**0.26**	**0.35**			
B	0.45	0.50	0.62			
C	**0.30**	**0.32**	0.50			
D	0.55	0.62	0.72			

Note: Combinations in the targeted toxicity interval (0.20-0.40) are in boldface. Schedules A, B, C, and D have dosing frequency of 192, 96, 48, and 24 hours, respectively.

violation of the monotonicity assumption of DLT probabilities with increasing exposure in Scenario 5. Moreover, Scenarios 1 to 5 assume completely ordered schedules, that is, DLT probabilities increase monotonically with schedules involving more frequent administration given the same dose. By contrast, Scenarios 6 and 7 relax this assumption, and assume partially ordered schedules. Scenario 6 correspond to Scenario 1 with DLT probabilities for Schedules B and C switched. Similarly, Scenario 7 corresponds to a scenario which spread across different intervals, but DLT probabilities for Schedules B and C are switched. In addition, we considered more scenarios to assess the performance of TITE-PK, which are listed in Table B1.

The elimination half-life of Vidaza is reported as 4 hours.[25] Thus, we specify the elimination rate constant $k_e = \frac{\log(2)}{4}$ 1/hours. In order to choose a meaningful value for the PK parameter k_{eff}, we sought for a sensible prior to model the parameter as random. A log-normal distribution constrains k_{eff} to positive values and allows for specification of a plausible range of values for the parameter. Consequently, we chose a log-normal distribution defined by the 0.025 and 0.975 quantiles reflecting the fastest and slowest time-scales of the experiment which are the absorption half-life and the cycle duration. It is reported that the Vidaza has a rapid absorption after subcutaneous administration,[25] which we interpret as a small value for the absorption rate, say 2 1/h. A log-normal distribution is specified by matching the inverse of cycle length 1/672 1/h and the absorption rate 2 1/h to the 0.025 and 0.975 quantiles, respectively. We can calculate the mean of the corresponding log-normal distribution by using the relationship between log-normal and normal distributions. Accordingly, the mean of the corresponding normal distribution μ is given by $\frac{\log(1/672)+\log(2)}{2} \approx -2.91$. The standard deviation of the corresponding normal distribution σ is given by $\frac{\log(2)-\log(1/672)}{2\times 1.96} \approx 1.84$. Then, the mean of the derived log-normal distribution can be calculated with the formula $\exp(\mu + \frac{\sigma^2}{2})$. Hence, the PK parameter k_{eff} is given by the mean of the log-normal distribution, that is, $k_{\text{eff}} = 0.295$. The calculated $\text{E}(t|d,f)$ and $\text{AUC}_E(t|d,f)$ of the Vidaza trial for combinations of the 8 mg-Schedule D and the 24 mg-Schedule B are displayed in Figure 2. Notice that the $\text{AUC}_E(t|d,f)$ of the 24

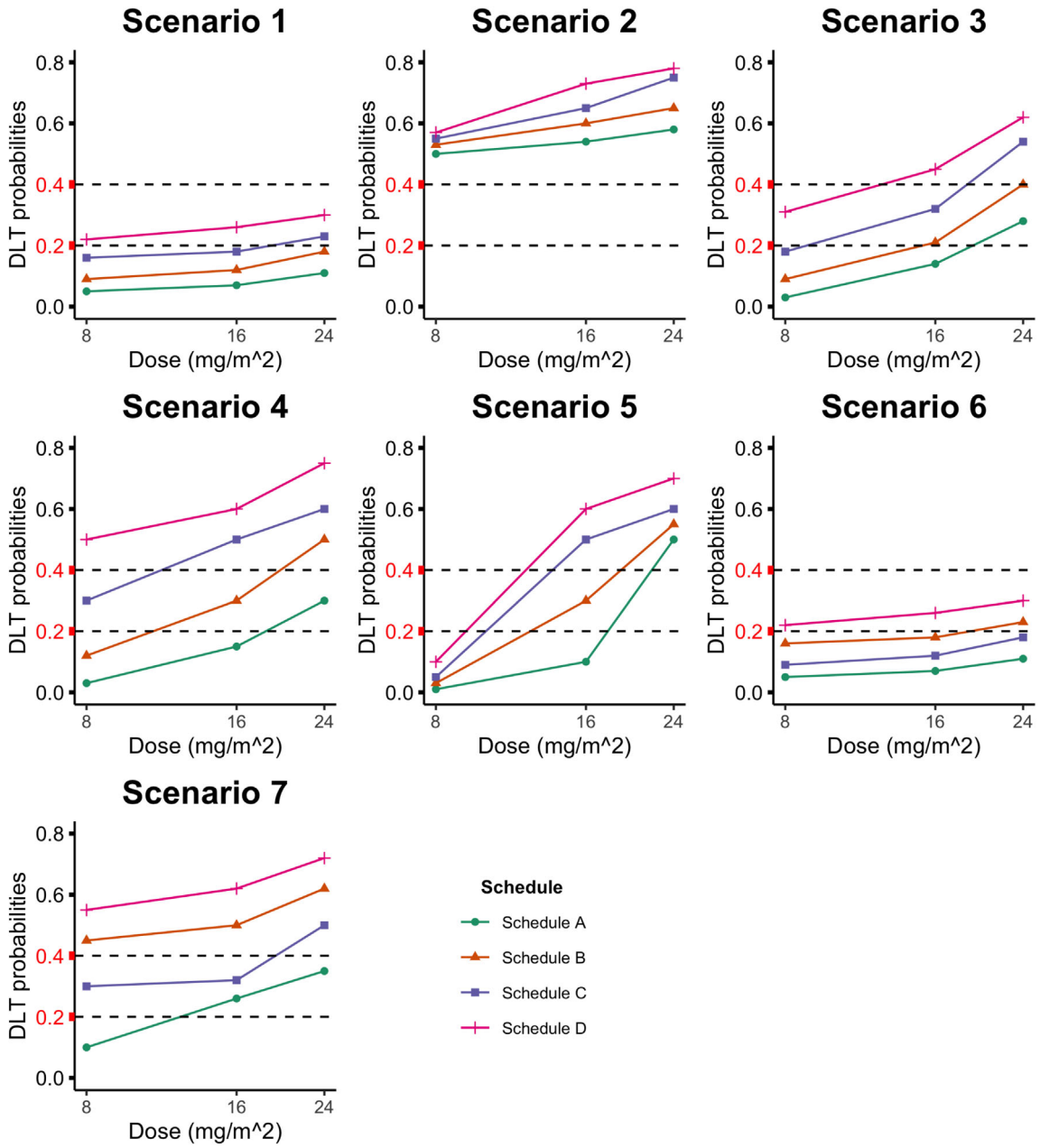

FIGURE 1 Toxicity scenarios for the dose-schedule combination in the simulation study. The horizontal dashed lines represent the boundaries of the targeted toxicity interval. Schedules A, B, C, and D have dosing frequency of 192, 96, 48, and 24 hours, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

mg-Schedule B at week 4 is 1, since this dose-schedule combination is taken as the reference dose-schedule combination and the length of cycle 1 is 4 weeks.

For TITE-PK model, a normal weakly informative prior is chosen for $\log(\beta)$ with a standard deviation of 1.75 and a mean which corresponds to $P(T \leq t^* | d^*, f^*)$ of 0.3. We used true values of k_e and k_{eff} for the estimation. The comparison of the prior DLT probabilities used by TITE-PK and the prior skeletons used by POCRM are displayed in Figure A1 in Appendix A.

We did not consider the method by Braun et al[5] in the simulations, since this method mostly requires a mean of approximately 60 patients to be enrolled which is not practical for many phase I trials. POCRM assumes that DLT probabilities increase monotonically with dose within each schedule. For different schedules, it specifies multiple possible orderings of dose-schedule combinations and uses model selection techniques to select the most appropriate model.[4] POCRM with partially ordered schedules relaxes the assumption of completely ordered schedules. This is done by specifying

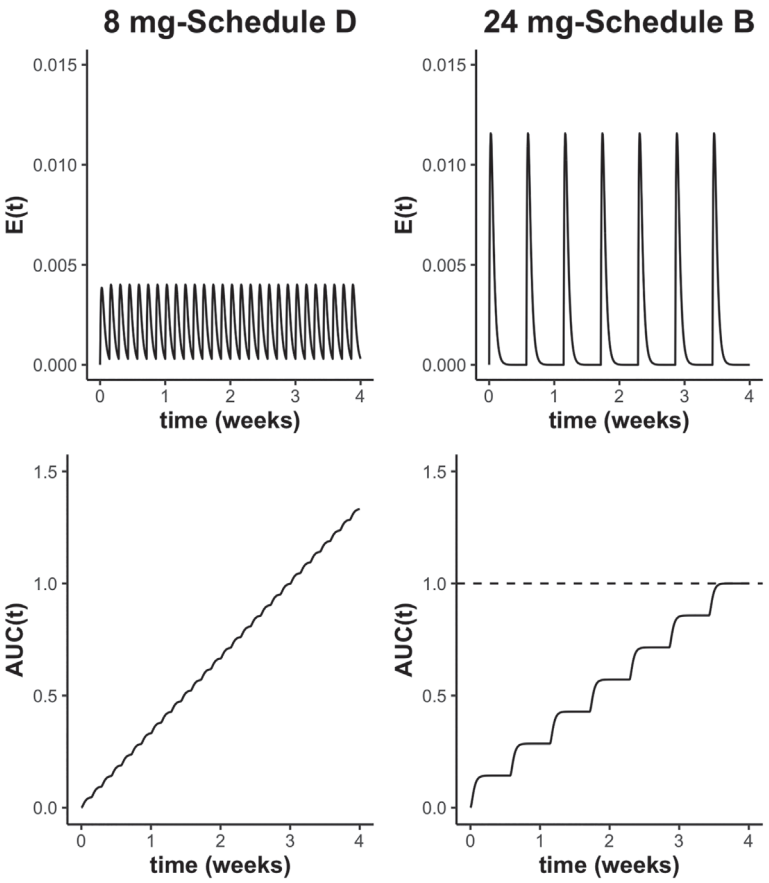

FIGURE 2 Illustration of the exposure measure of the drug (E($t|d,f$)) and the AUC over exposure measure (AUC$_E$($t|d,f$)) for the 8 mg-Schedule D and the 24 mg-Schedule B combinations of the Vidaza trial, respectively. The reference dose-schedule combination is the 24 mg-Schedule B and the length of cycle 1 is 4 weeks

appropriate possible orderings. In the simulations, two versions of the POCRM, one assuming complete order schedules and one with partial order schedules, are considered. Computations of the POCRM were carried out using the publicly available R code provided by Wages et al[7] (see http://faculty.virginia.edu/model-based_dose-finding). We refer to Wages et al[7] for more information about the POCRM.

In the simulations, data for 1000 trials were generated per scenario. For all methods, patients were assigned one at a time until the trial was stopped, the MTC was identified or the maximum number of patients per trial of 60 were used. For the TITE-PK, if all dose-schedule combinations are in the OD interval based on the adapted EWOC criterion, the trial is stopped without selecting any combination as the MTC. Otherwise, the trial continues until the recommendation of the MTC. The recommended MTC must meet the following conditions:

(i) At least nine patients have been treated at the MTC.
(ii) The recommended MTC satisfies one of the following conditions:
- The probability of TT at the MTC exceeds 50%: $P(0.16 \leq P(T \leq t^*|d,f) \leq 0.33) \geq 0.50$.
- A minimum of 21 patients have already been treated in the trial.

We consider two values for the feasibility bound, namely, $a = 0.25$ and $a = 0.50$. Increasing the feasibility bound results in the more aggressive dose-schedule decisions, that is, increasing the percentage of trials with the MTC declared in the TT interval, while increasing the percentage of trials with the MTC declared in the OD interval.

3.2 | Results

Table 2 shows the summary statistics for the performance of the methods under the seven scenarios. In Scenario 1, TITE-PK with $a = 0.50$ outperforms both POCRM methods in terms of recommending the MTC in the TT interval. The corresponding percentages are 76%, 62%, and 65% for TITE-PK ($a = 0.50$), POCRM (complete), and POCRM (partial), respectively. However, TITE-PK with $a = 0.25$ yields the worst performance according to the same measure, the

TABLE 2 Simulation results for two applications of POCRM with complete and partial order schedules and two applications of the proposed method TITE-PK with a feasibility bound of $a = 0.25$ and $a = 0.50$

	Scenario						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Probability of selecting MTC in the targeted toxicity interval							
POCRM (complete)	0.62	n/a	0.74	0.65	0.42	0.57	0.52
POCRM (partial)	0.65	n/a	0.74	0.63	0.44	0.66	0.57
TITE-PK (a = 0.25)	0.44	n/a	0.58	0.38	0.25	0.39	0.31
TITE-PK (a = 0.50)	0.76	n/a	0.79	0.60	0.40	0.76	0.57
Probability of selecting MTC in the overdosing interval							
POCRM (complete)	0.00	0.49	0.08	0.20	0.36	0.00	0.26
POCRM (partial)	0.00	0.50	0.10	0.21	0.36	0.00	0.26
TITE-PK (a = 0.25)	0.00	0.07	0.01	0.03	0.04	0.00	0.04
TITE-PK (a = 0.50)	0.00	0.28	0.10	0.14	0.20	0.00	0.24
Probability of selecting no combination as MTC							
POCRM (complete)	0.05	0.51	0.03	0.03	0.01	0.06	0.09
POCRM (partial)	0.05	0.50	0.03	0.03	0.01	0.02	0.10
TITE-PK (a = 0.25)	0.10	0.93	0.13	0.15	0.07	0.11	0.37
TITE-PK (a = 0.50)	0.04	0.72	0.03	0.03	0.01	0.05	0.08
Mean number of patients enrolled in the overdosing interval							
POCRM (complete)	0.0	17.0	2.5	6.1	9.3	0.0	8.1
POCRM (partial)	0.0	17.6	3.2	7.0	10.1	0.0	8.2
TITE-PK (a = 0.25)	0.0	3.8	1.2	2.7	5.1	0.0	4.0
TITE-PK (a = 0.50)	0.0	8.7	4.5	7.1	10.1	0.0	8.3
Mean number of patients enrolled in total							
POCRM (complete)	25.6	17.0	24.5	24.2	24.6	25.2	22.2
POCRM (partial)	25.9	17.6	25.7	25.3	25.8	25.9	23.6
TITE-PK (a = 0.25)	21.5	3.8	19.1	18.6	19.0	21.1	14.2
TITE-PK (a = 0.50)	18.8	8.7	20.5	20.8	21.7	18.3	19.9
Mean number of DLT observed							
POCRM (complete)	4.6	8.9	6.6	7.2	7.1	4.5	7.6
POCRM (partial)	4.7	9.3	6.9	7.8	7.7	4.7	8.3
TITE-PK (a = 0.25)	3.8	2.0	4.6	4.7	4.8	3.8	4.3
TITE-PK (a = 0.50)	4.2	4.9	6.6	7.1	7.6	4.1	7.6

Abbreviations: MTC, maximum tolerated dose-schedule combination; POCRM, partial order continual reassessment method; TITE-PK, time-to-event pharmacokinetic.

corresponding percentage is 44%. Scenario 1 includes no dose-schedule combinations in the UD interval, hence probability of selecting no combination as the MTC is 0 for all methods. In Scenario 2, all combinations are in the OD interval. TITE-PK with $a = 0.25$ and $a = 0.50$ stop the trial without the MTC selection in 93% and 72% of the time, respectively. However, both POCRM methods stop the trial around 50% of the time. Subsequently, the percentages of the MTC selection in the OD interval of POCRM methods are higher than both TITE-PK models. Moreover, TITE-PK with $a = 0.50$ is superior to POCRM with partial ordering in terms of mean number of patients enrolled in the OD interval (8.7 vs 17.6) and mean number of DLT observed (4.9 vs 9.3). An advantage of TITE-PK is that, due to early stopping and a small mean number of patients, the design expose approximately half of the sample size to toxic combinations in comparison to POCRM with partial order (given above).

In Scenario 3, TITE-PK with $a = 0.50$ gives higher percentage for recommending a combination in the TT interval than both POCRM methods. The corresponding percentages are 79% for TITE-PK ($a = 0.50$), 74% for POCRM (complete), and 74% for POCRM (partial). TITE-PK with $a = 0.25$ selects the MTC in the OD interval in about 1% of the time for Scenario 3, while it yields the worst performance with 58% in terms of recommending a combination in the TT interval. In Scenario 4, both POCRM methods yield higher percentages than TITE-PK with $a = 0.50$ for the selection of the MTC in the TT interval. The corresponding percentages are 60%, 65%, and 63% for TITE-PK ($a = 0.50$), POCRM (complete), and POCRM (partial), respectively. However, both POCRM methods yield slightly higher percentages than TITE-PK with $a = 0.50$ for the selection of the MTC in the OD interval. TITE-PK with $a = 0.50$ recommends the MTC in the OD interval in 14% of the trials, while POCRM with complete ordering and POCRM with partial ordering do this in 20% and 21% of the trials.

Scenario 5 needs special consideration, since all methods perform poorly in terms of the MTC selection in the TT interval. Consistent with other scenarios, TITE-PK ($a = 0.25$) results in the lowest probability for the MTC selection in the TT interval. However, both TITE-PK methods display superior performance in terms of selecting MTC in the OD interval. The corresponding percentages are 20% for TITE-PK ($a = 0.50$) and 36% for POCRM with partial ordering. Scenario 6 corresponds to Scenario 1 but with DLT probabilities for Schedules B and C switched. In Scenario 6, TITE-PK ($a = 0.50$) displays better performance compared with POCRM methods as in Scenario 1. In Scenario 7, TITE-PK ($a = 0.50$) yield lower percentages than POCRM methods in terms of selecting the MTC in the OD interval. Although Scenarios 6 and 7 assume partial orderings, our method performs relatively well showing robustness against the violation of the complete ordering assumption.

We also examined which schedules are recommended by TITE-PK ($a = 0.50$) and POCRM with partial ordering as part of the MTC. These results are listed in Table 3. In many scenarios, the schedules recommended by POCRM are more spread across four schedules compared with the schedules recommended by TITE-PK. For example, in Scenario 5 the MTC selection in the TT interval is similar for two methods (the only combination in the TT interval is in Schedule B). However, the selected schedules by two methods are quite different. This results in inferior performance of POCRM in terms of the MTC selection in the OD interval. This is because Schedule A of Scenario 5 which was selected by TITE-PK in 45% of the time has less toxic dose-schedule combinations than Schedules C and D. The fact that TITE-PK selects the schedules more precisely is reflected in its superior performance in terms of the MTC selection in the TT and/or OD intervals. One possible reason is that the EWOC criterion used by TITE-PK does not allow to escalate to the schedules with higher toxicity in comparison to POCRM, hence improving the overall performance.

Overall, TITE-PK with $a = 0.25$ yields more conservative behavior in terms of the MTC selection in the TT and the OD intervals in comparison to TITE-PK with $a = 0.50$, as it is expected. In all scenarios, TITE-PK with $a = 0.25$ does not select

TABLE 3 Simulation results for schedules recommended by TITE-PK (a = 0.50) and POCRM (partial) as part of the MTC

	Scenario						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Probability of selecting Schedule A as part of MTC							
TITE-PK	0.00	0.27	0.10	0.31	0.45	0.01	0.52
POCRM	0.02	0.45	0.17	0.28	0.16	0.02	0.36
Probability of selecting Schedule B as part of MTC							
TITE-PK	0.06	0.02	0.47	0.59	0.47	0.07	0.19
POCRM	0.19	0.04	0.30	0.34	0.45	0.27	0.20
Probability of selecting Schedule C as part of MTC							
TITE-PK	0.22	0.00	0.34	0.07	0.08	0.13	0.19
POCRM	0.23	0.01	0.29	0.29	0.27	0.15	0.33
Probability of selecting Schedule D as part of MTC							
TITE-PK	0.67	0.00	0.06	0.00	0.00	0.74	0.01
POCRM	0.51	0.00	0.20	0.06	0.10	0.50	0.02

Note: Schedules A, B, C, and D have dosing frequency of 192, 96, 48, and 24 hours, respectively.
Abbreviations: MTC, maximum tolerated dose-schedule combination; POCRM, partial order continual reassessment method; TITE-PK, time-to-event pharmacokinetic.

TABLE 4 Simulation results under different time-to-DLT distributions: TITE-PK, uniform and exponential distributions, and time-to-DLT occurring with higher probability at the early (between time 0 and $\frac{t^*}{5}$) or late (between time $\frac{4\ t^*}{5}$ and t^*) stage within the first cycle

	Scenario						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Probability of selecting MTC in the targeted toxicity interval							
TITE-PK	0.76	0.00	0.79	0.60	0.40	0.76	0.57
Uniform	0.77	0.00	0.80	0.59	0.45	0.78	0.55
Exponential	0.74	0.00	0.79	0.57	0.45	0.76	0.54
Early/late	0.74	0.00	0.80	0.57	0.44	0.76	0.53
Mean number of patients enrolled in the overdosing interval							
TITE-PK	0.0	8.7	4.5	7.1	10.1	0.0	8.3
Uniform	0.0	9.9	4.8	8.2	10.3	0.0	8.6
Exponential	0.0	9.3	4.9	7.7	10.0	0.0	8.7
Early/late	0.0	9.4	5.2	8.3	9.8	0.0	8.7
Mean number of patients enrolled in total							
TITE-PK	18.8	8.7	20.5	20.8	21.7	18.3	19.9
Uniform	18.5	9.9	20.4	21.0	21.2	18.4	19.7
Exponential	18.6	9.3	20.3	20.9	21.3	18.4	19.9
Early/late	19.2	9.4	20.4	20.9	21.0	18.7	19.4

Abbreviations: DLT, dose-limiting toxicity; MTC, maximum tolerated dose-schedule combination; TITE-PK, time-to-event pharmacokinetic.

the MTC in the OD interval more than 7% of the time. Furthermore, TITE-PK with $a = 0.25$ induces the lowest number of DLT in all scenarios and enroll the lowest number of patients to the OD interval. However, this conservative behavior consistently results in a weaker performance in terms of selecting the MTC in the TT interval. The main reason of this poor behavior is related to the EWOC criterion and the choice of $a = 0.25$. TITE-PK with $a = 0.50$ yields superior or similar performance compared with the POCRM methods in terms of selecting the MTC in the TT interval with the exception of Scenario 4. In terms of the MTC selection in the OD interval, TITE-PK with $a = 0.50$ performs consistently better than POCRM methods. Furthermore, POCRM (partial) does not display clear benefit over TITE-PK with $a = 0.50$ for Scenario 6 and 7 in which the assumption of complete ordering is relaxed. Finally, both TITE-PK models enroll lower number of patients compared with both POCRM methods in all scenarios. One reason of the desirable performance of TITE-PK may stem from the use of EWOC criterion, which reduces the risk of recommending toxic dose-schedule combinations as the MTC.

With the proposed method TITE-PK, time-to-DLT is modeled using a nonhomogeneous Poisson distribution which has been used to simulate the timing of the events in the previously discussed scenarios. To examine the robustness of the Poisson process assumption following the exposure metric in TITE-PK, we generated datasets from different time-to-DLT models, namely, uniform and exponential distributions, under each scenario. Furthermore, we considered a third data-generating process, that is, assuming time-to-DLT occurring with higher probability at the early (between 0 and $\frac{t^*}{5}$) or late (between time $\frac{4\ t^*}{5}$ and t^*) stages within the first cycle. For the uniform distribution, the occurrences of DLT are determined using the true DLT probabilities. The timing of DLT is sampled uniformly within the first cycle. For the exponential distribution, the rate parameter of the exponential distribution is calculated by $\lambda = -\log(1 - P(T \leq t^*|d,f))/t^*$ where $P(T \leq t^*|d,f)$ corresponds to a true DLT probability.[26] Then, the timing of DLT is sampled using the exponential distribution with the specified rate parameter within the first cycle. For the third data-generating process, the occurrences of DLT are determined using the true DLT probabilities as in the uniform distribution. The timing of DLT is sampled assuming with the probability of 0.4 for the interval 0 and $\frac{t^*}{5}$ (early) and the probability of 0.4 for the interval $\frac{4\ t^*}{5}$ and t^* (late). These give that within the interval $\frac{t^*}{5}$ and $\frac{4\ t^*}{5}$, the corresponding probability is 0.2. We only considered the feasibility bound of $a = 0.50$. Table 4 gives the results of three performance measures. The first rows of each performance measure replicate the values displayed in Table 2 of TITE-PK ($a = 0.50$) values. Table 4 indicates that the performance of TITE-PK varies little with the time-to-DLT distribution in terms of investigated measures.

4 | DISCUSSION AND CONCLUSIONS

We propose a Bayesian adaptive model, a TITE-PK model, to support design and analysis of phase I dose-escalation trials with multiple schedules to provide guidance for the dose-schedule decisions. TITE-PK has an interpretable parameter which facilitates the prior specification. It uses PK principles to combine different treatment schedules in a model-based approach. An adapted EWOC criterion can be used with TITE-PK. In the simulations, for six of seven scenarios considered, TITE-PK with a feasibility bound of 0.50 shows superior or similar performance in terms of identifying the MTC in the TT interval compared with the POCRM. In terms of the recommendation of the MTC in the OD interval, TITE-PK yields lower percentages in all seven scenarios considered. For all scenarios, the TITE-PK model required lower numbers of patients enrolled compared with POCRM.

Here, we considered simultaneously finding a suitable dose-schedule combination within a phase I trial as in Wages et al[7] Another useful design would investigate multiple schedules, say Schedules 1 and 2, sequentially. That is, enrolling dose cohorts of patients with Schedule 1 and estimating the maximum tolerated dose (MTD) for the Schedule 1. Then, patients are enrolled into dose cohorts using Schedule 2 and the MTD is estimated for Schedule 2 by utilizing data coming from both schedules. TITE-PK can be used to design such sequential phase I trials or a phase I trial involving only one schedule. As such designs are beyond the scope of this article, they are not investigated here.

Here, we considered the cohort size of 1. However, there is no restriction in TITE-PK regarding the cohort size. We defined the schedule as the frequency of administration within a cycle. Nevertheless, TITE-PK can be used to design trials with the other definition of the schedule, that is, the number of cycles to treat patients. This can be achieved by assigning different reference time point t^* for different dose-schedule combinations based on the number of treatment cycles. Frequency of administrations f for different dose-schedule combinations will be assumed to be the same.

One limitation of the proposed method is the assumption of monotonicity of the exposure-DLT probability relationship. Moreover, this monotonicity assumption implies a monotonic dose-DLT curve, since a linear PK is used in the pseudo-PK model. Violation of the linear PK assumption can be informed using the external PK data from the ongoing trial. To relax the assumption of linear PK, one can consider more complicated PK models including a nonlinear PK model which may not have an analytical solution. Such extensions may be implemented in `Stan` which has a built-in differential equation solver. However, more complicated modeling approaches always need to be calibrated well given the sparseness of the phase I dose-escalation datasets. Alternatively, one can consider an ad hoc extension of the TITE-PK model. For instance, by introducing a nonlinearity factor γ,[27] a *pseudo*-dose as $(\frac{d}{d^*})^\gamma$ can be used instead of dose d in the model which may be helpful to relax the linear PK assumption.

When relevant historical information or data from a different study population exists, it is desirable to include such information in the analysis of the phase I trial, for example, using a meta-analytic-predictive (MAP) prior.[28] TITE-PK can be extended to use a MAP approach. Another crucial aspect of the methods for phase I trials is the ability to analyse the combination of drugs. Although we only consider the single agent case here, it is possible to extend TITE-PK to analyse drug combinations which is complicated by the need to model possible drug interactions. Another extension of TITE-PK is considering a two-parameter version in which one of the PK parameters k_{eff} is also estimated in the model jointly with the regression coefficient β.

ACKNOWLEDGEMENTS

We thank Heinz Schmidli who pointed us to several important references, Michael Looby for proofreading an earlier version of the article, Abdelkader Seroutou and Christian Röver for contributing valuable comments. We thank the Associate Editor and two anonymous reviewers whose comments and suggestions helped to improve this article.

ORCID

Burak Kürsad Günhan https://orcid.org/0000-0002-7454-8680
Tim Friede https://orcid.org/0000-0001-5347-7441

REFERENCES

1. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*. 1990;46(1):33-48.
2. Neuenschwander B, Matano A, Tang Z, Roychoudhury S, Wandel S, Bailey SA. Bayesian industry approach to phase I combination trials in oncology. *Statistical Methods in Drug Combination Studies*. Boca Raton, FL: CRC Press; 2015:95-135.

3. Guo B, Li Y, Yuan Y. A dose-schedule finding design for phase I-II clinical trials. *J R Stat Soc C*. 2016;65(2):259-272.
4. O'Quigley J, Iasonos A, Bornkamp B. *Handbook of Methods for Designing, Monitoring, and Analyzing Dose-Finding Trials*. London: Routledge; 2017:127-139.
5. Braun T, Thall PF, Nguyen H, De Lima M. Simultaneously optimizing dose and schedule of a new cytotoxic agent. *Clin Trials*. 2007;4(2):113-124.
6. Zhang J, Braun T. A phase I Bayesian adaptive design to simultaneously optimize dose and schedule assignments both between and within patients. *J Am Stat Assoc*. 2013;108(503):892-901.
7. Wages N, O'Quigley J, Conaway M. Phase I design for completely or partially ordered treatment schedules. *Stat Med*. 2014;33(4):569-579. https://doi.org/10.1002/sim.5998.
8. Li Y, Bekele B, Ji Y, Cook J. Dose-schedule finding in phase I/II clinical trials using a Bayesian isotonic transformation. *Stat Med*. 2008;27(24):4895-4913.
9. Thall P, Nguyen H, Braun T, Qazilbash M. Using joint utilities of the times to response and toxicity to adaptively optimize schedule-dose regimes. *Biometrics*. 2013;69(3):673-682.
10. Cunanan K, Koopmeiners J. A Bayesian adaptive phase I-II trial design for optimizing the schedule of therapeutic cancer vaccines. *Stat Med*. 2017;36(1):43-53.
11. Pinheiro J, Duffull S. Exposure response – getting the dose right. *Pharm Stat*. 2009;8(3):173-175. https://doi.org/10.1002/pst.401.
12. Cox E, Veyrat-Follet C, Beal S, Fuseau E, Kenkare S, Sheiner L. A population pharmacokinetic–pharmacodynamic analysis of repeated measures time-to-event pharmacodynamic responses: the antiemetic effect of ondansetron. *J Pharmacokinet Biopharm*. 1999;27(6):625-644.
13. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Stat Med*. 1998;17(10):1103-1120. https://doi.org/10.1002/(SICI)1097-0258(19980530)17:10<1103::AID-SIM793>3.0.CO;2-9.
14. Ursino M, Zohar S, Lentz F, et al. Dose-finding methods for phase I clinical trials using pharmacokinetics in small populations. *Biom J*. 2017;59(4):804-825. https://doi.org/10.1002/bimj.201600084.
15. Jacqmin P, Snoeck E, Van Schaick E, et al. Modelling response time profiles in the absence of drug concentrations: definition and performance evaluation of the K-PD model. *J Pharmacokinet Pharmacodyn*. 2007;34(1):57-85. https://doi.org/10.1007/s10928-006-9035-z.
16. Jacobs T, Straetemans R, Molenberghs G, Adriaan BJ, Bijnens L. A latent pharmacokinetic time profile to model dose-response survival data. *J Biopharm Stat*. 2010;20(4):759-767.
17. Wages N, Conaway M, O'Quigley J. Dose-finding design for multi-drug combinations. *Clin Trials*. 2011;8(4):380-389.
18. De Lima M, Giralt S, Thall P, et al. Maintenance therapy with low-dose azacitidine after allogeneic hematopoietic stem cell transplantation for recurrent acute myelogenous leukemia or myelodysplastic syndrome. *Cancer*. 2010;116(23):5420-5431.
19. Kalbfleisch J, Prentice RL. *The Statistical Analysis of Failure Time Data*. New York, NY: John Wiley & Sons; 2002:31-52.
20. Cheung Y, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*. 2000;56(4):1177-1182.
21. Kallen A. *Computational Pharmacokinetics*. Boca Raton, FL: CRC Press; 2007:13-28.
22. Bertrand, J and Mentré, F. Mathematical expressions of the pharmacokinetic and pharmacodynamic models implemented in the Monolix software; 2008. lixoft.com/wp-content/uploads/2016/03/PKPDlibrary.pdf. Accessed September, 2018.
23. Carpenter B, Gelman A, Hoffman M, et al. Stan: a probabilistic programming language. *J Stat Softw*. 2017;76(1):1-32.
24. Gelman A, Rubin D. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7(4):457-472.
25. Vidaza [package insert] U.S. food and drug administration; 2018. https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/050794s031lbl.pdf. Accessed June, 2019.
26. Mood A, Graybill F, Boes DS. *Introduction to the Theory of statistics*. New York, NY: McGraw-Hill; 1974:85-129.
27. O'Quigley J, Chevret S. Methods for dose finding studies in cancer clinical trials: a review and results of a Monte Carlo study. *Stat Med*. 1991;10(11):1647-1664.
28. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*. 2014;70(4):1023-1032. https://doi.org/10.1111/biom.12242.

How to cite this article: Günhan BK, Weber S, Friede T. A Bayesian time-to-event pharmacokinetic model for phase I dose-escalation trials with multiple schedules. *Statistics in Medicine*. 2020;39:3986–4000. https://doi.org/10.1002/sim.8703

APPENDIX A. PRIOR PROBABILITIES FOR TITE-PK AND PRIOR SKELETONS FOR POCRM

Figure A1 shows the comparison of the prior DLT probabilities used by TITE-PK and the prior skeletons used by POCRM.

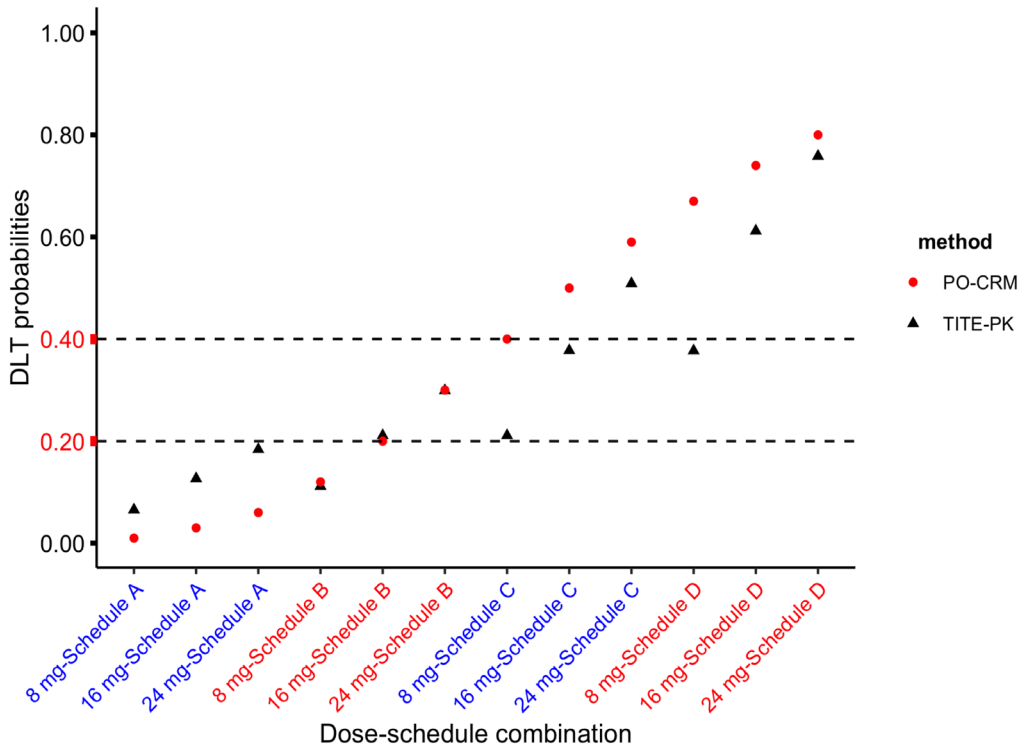

FIGURE A1 Prior DLT probabilities obtained by TITE-PK (as triangles) and the prior skeletons used by POCRM (as circles). The horizontal dashed lines represents the boundaries of the targeted toxicity interval. Schedules A, B, C, and D have dosing frequency of 192, 96, 48, and 24 hours, respectively. DLT, dose-limiting toxicity; POCRM, partial order continual reassessment method; TITE-PK, time-to-event pharmacokinetic [Colour figure can be viewed at wileyonlinelibrary.com]

APPENDIX B. ADDITIONAL SIMULATION RESULTS

We also conducted simulations to investigate further toxicity scenarios. The scenarios and the results are shown in Tables B1 and B2, respectively.

TABLE B1 Additional simulation scenarios: Toxicity scenarios for the dose-schedule combination in the simulation study

	Doses in mg/m^2					
Schedule	**8**	**16**	**24**	**8**	**16**	**24**
	Scenario 8			Scenario 9		
A	0.10	0.26	0.35	0.10	0.28	0.45
B	0.30	0.32	0.50	0.12	**0.30**	0.48
C	0.45	0.50	0.62	0.14	**0.32**	0.55
D	0.55	0.62	0.72	**0.30**	0.48	0.70
	Scenario 10					
A	0.01	0.10	0.50			
B	0.05	0.50	0.60			
C	0.03	**0.30**	0.55			
D	0.10	0.60	0.70			

Note: Combinations in the targeted toxicity interval (0.20-0.40) are in boldface. Schedules A, B, C, and D have dosing frequency of 192, 96, 48, and 24 hours, respectively.

TABLE B2 Additional simulation results of POCRM with partial order schedules and the proposed method TITE-PK with a feasibility bound of $a = 0.50$

	Scenario		
	8	**9**	**10**
Probability of selecting MTC in the targeted toxicity interval			
POCRM (partial)	0.62	0.63	0.44
TITE-PK (a = 0.50)	0.69	0.75	0.22
Probability of selecting MTC in the overdosing interval			
POCRM (partial)	0.22	0.17	0.33
TITE-PK (a = 0.50)	0.12	0.13	0.31
Probability of selecting no combination as MTC			
POCRM (partial)	0.10	0.10	0.01
TITE-PK (a = 0.25)	0.11	0.08	0.01
Mean number of patients enrolled in the overdosing interval			
POCRM (partial)	7.9	5.7	10.0
TITE-PK (a = 0.50)	6.5	7.0	11.5
Mean number of patients enrolled in total			
POCRM (partial)	24.0	24.8	25.7
TITE-PK (a = 0.50)	19.5	19.5	21.9
Mean number of DLT observed			
POCRM (partial)	8.4	7.4	7.7
TITE-PK (a = 0.50)	7.1	7.1	8.0

Abbreviations: DLT, dose-limiting toxicity; MTC, maximum tolerated dose-schedule combination; POCRM, partial order continual reassessment method; TITE-PK, time-to-event pharmacokinetic.

A.2 Sequential phase I dose-escalation trials with multiple schedules

The paper is currently under review. Preprint version (20.08.2020) is publicly available from https://arxiv.org/abs/1811.09433.

A Bayesian time-to-event pharmacokinetic model for sequential phase I dose-escalation trials with multiple schedules

Burak Kürsad Günhan,[1 2] **Sebastian Weber**,[3] **Abdelkader Seroutou**,[3] **Tim Friede**[1]

Phase I dose-escalation trials constitute the first step in investigating the safety of potentially promising drugs in humans. Conventional methods for phase I dose-escalation trials are based on a single treatment schedule only. More recently, however, multiple schedules are more frequently investigated in the same trial. Here, we consider sequential phase I trials, where the trial proceeds with a new schedule (e.g. daily or weekly dosing) once the dose escalation with another schedule has been completed. The aim is to utilize the information from both the completed and the ongoing dose-escalation trial to inform decisions on the dose level for the next dose cohort. For this purpose, we adapted the time-to-event pharmacokinetics (TITE-PK) model, which were originally developed for simultaneous investigation of multiple schedules. TITE-PK integrates information from multiple schedules using a pharmacokinetics (PK) model. In a simulation study, the developed appraoch is compared to the bridging continual reassessment method and the Bayesian logistic regression model using a meta-analytic-prior. TITE-PK results in better performance than comparators in terms of recommending acceptable dose and avoiding overly toxic doses for sequential phase I trials in most of the scenarios considered. Furthermore, better performance of TITE-PK is achieved while requiring similar number of patients in the simulated trials. For the scenarios involving one schedule, TITE-PK displays similar performance with alternatives in terms of acceptable dose recommendations. The `R` and `Stan` code for the implementation of an illustrative sequential phase I trial example is publicly available (https://github.com/gunhanb/TITEPK_sequential). In sequential phase I dose-escalation trials, the use of all relevant information is of great importance. For these trials, the adapted TITE-PK which combines information using PK principles is recommended.

Keywords: Phase I dose-escalation trials, multiple treatment schedules, PK models, Bayesian statistics

1 Background

Phase I dose-escalation trials constitute the first step in investigating the safety of potentially promising drugs in humans[1]. In oncology, such trials focus on identifying the maximum tolerated dose (MTD) through a series of dose-escalation steps. Dose-escalation trials traditionally enroll small cohorts of patients who are treated in cycles. Typically, the estimation of the MTD is based on the toxicity data of the first cycle only. The observed toxicities are classified into dose-limiting toxicities (DLT) and non-DLT. Each time a cohort completes the first cycle at a given dose level, the available data are assessed to decide how the trial proceeds. A commonly accepted target for the MTD in oncology is to allow for a DLT probability of 33% per cycle of treatment.

Standard statistical methods include adaptive model based approaches such as the continual reassessment method (CRM)[2] or the Bayesian logistic regression model (BLRM)[3]. The BLRM is a two-parameter version of the CRM which utilizes the escalation with the overdose control (EWOC)[4]

[1] *Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany*
[2] *Correspondence to: Burak Kürsad Günhan; email:* `burak.gunhan@med.uni-goettingen.de`
[3] *Novartis Pharma AG, Basel, Switzerland*

criterion. The EWOC criterion aims to reduce the risk of overdosing patients by choosing doses with a posterior probability of being above the true MTD lower than a feasibility bound.

In addition to the dose administered, the frequency of administration, known as the schedule, is a crucial part of a treatment plan of any phase I trial. In practice, sometimes it is required to investigate multiple schedules, e.g. a dose given once a day or an adequately larger dose given once a week. Hence, the probability of DLT for each patient is a function of both the dose and the schedule. Simultaneous investigation of dose and schedule within a phase I trial has gained some attention in the literature. In such trials, the doses and the schedules are altered for different cohorts of patients within the same trial. Methods for simultaneous investigation of dose and schedule combination include a Bayesian time-to-event model by Braun et al[5] and the partial order continual reassessment method by Wages et al[6]. Recently, Günhan et al[7] proposed an alternative dose-schedule finding method, a Bayesian time-to-event pharmacokinetics model (TITE-PK), which uses pharmacokinetics (PK) principles. Unlike other phase I methods, TITE-PK makes use of an exposure-response model that is often more informative than a standard dose-response model. TITE-PK models the relationship between time-to-first DLT and an exposure measure of the drug obtained by a *pseudo*-PK model in a Bayesian model-based approach. TITE-PK has been shown to have desirable operating characteristics in terms of finding an acceptable dose and schedule simultaneously in simulation studies[7].

In this paper, we consider an alternative phase I design in which multiple treatment schedules are investigated sequentially, rather than simultaneously. The schedules are denoted by S_i where $i = 1, 2, \ldots, k$. The sequential multiple schedule design proceeds as follows. In the first step, cohorts of patients are enrolled with S_1 and the trial is continued until the MTD is declared for S_1. In the second step, the trial continues with schedule S_2 and the starting dose can be informed from the S_1. Dose-escalation decisions are informed by utilizing information from both schedules S_1 and S_2. That is, data from both the completed schedule S_1 and the ongoing schedule S_2 are integrated. Once the MTD for the Schedule S_2 is determined, the trial can continue with schedule S_3 and so on.

A sequential phase I trial with different strata, where strata may correspond to different patient populations, formulations, or treatment schedules etc., also called as a bridging trial, was considered by Liu et al[8] among others[9,10,11]. Liu et al[8] introduced the bridging CRM to borrow information from different strata. B-CRM takes into account potential heterogeneity between different strata using a Bayesian model averaging approach. Neuenschwander et al[9] suggest the use of BLRM with a meta-analytic-predictive (MAP) prior[12] approach (BLRM-MAP) to take advantage of the completed trial with different strata.

Borrowing approaches are based on discounting the existing information at the cost of increasing the needed saample size to achieve an acceptable performance in a new trial. Here we suggest the use of a modelling approach based on PK principles in order to increase the statistical efficiency. Therefore, we adapted the TITE-PK to design and analyze sequential phase I trials with multiple schedules. In the first step, TITE-PK is used to inform dose-escalation decisions for schedule S_1 until the MTD is declared or the trial is stopped. In the next step, TITE-PK models the data from both the completed (S_1) and the ongoing (S_2) trial directly, but only recommending doses for Schedule S_2. TITE-PK can be used for any number of schedules. We investigate the operating characteristics of TITE-PK for phase I trials with one schedule and sequential phase I trials with multiple schedules through simulations. We provide simulation results comparing the performance of TITE-PK to CRM and BLRM for phase I trials involving one schedule and to B-CRM and BLRM-MAP for sequential phase I trials involving multiple schedules.

This paper is organized as follows. In the following section, we describe an illustrative phase I trial example from oncology which investigated daily and weekly treatment schedules. Then, we describe the adapted TITE-PK for sequential investigation of multiple schedules. The performance

Table 1: Data of the everolimus trial. The treatment schedules which are used, the doses which are administered in mg/m^2, number of patients, and number of DLT are given.

Schedule	Dose (mg/m^2)	Number of patients	Number of DLT
Weekly	20.0	5	0
Weekly	30.0	13	4
Daily	2.5	4	2
Daily	5.0	6	3

of TITE-PK and comparators are studied in simulations. Later, different methods are applied to the illustrative example. We close with discussion and conclusions.

1.1 Illustrative example: Everolimus trial

Everolimus (RAD001) is an oral inhibitor of mammalian target of rapamycin, that has been developed as an antitumor agent[13]. Everolimus is approved by the US FDA to treat various conditions including certain types of pancreatic cancer and gastrointestinal cancer[13] and certain type of tuberous sclerosis[14]. The elimination half-life and the absorption rate of everolimus for cancer patients were reported as 30 (hours) and 2.5 (1/hours), respectively[15]. Everolimus was included in a phase Ib trial in combination with standard of care (etoposide and cisplatin chemotherapy) to identify a feasible dose and schedule in the treatment of small cell lung cancer (ClinicalTrials.gov identifier: NCT00466466)[16]. The trial was open-label and multi-centered. Patients were assigned alternately to either weekly or daily schedules of everolimus in treatment cycles of 21 days. In the everolimus trial, doses in both schedules were escalated simultaneously and analysed separately from one another. A Bayesian time-to-event model[17] was used to inform the dose-escalation decisions. The final data can be obtained from the supplementary material of Besse et al[16]. The dataset is displayed in Table 1. All DLT were reported at day 15. Based on investigator and medical monitor opinion, 2.5 mg/m^2 with daily schedule was identified as the MTD[16].

We used this trial to illustrate the TITE-PK approach for sequential designs, because (1) the trial evaluated two different schedules (weekly and daily dosing) and (2) the large number of DLT allows a good assessment on the performance of the TITE-PK. We will analyse the final dataset as if the trial had been conducted sequentially, specifically assuming S_1 is weekly schedule and S_2 is daily schedule.

2 Methods

2.1 TITE-PK for sequential phase I trials

TITE-PK for simultaneous investigation of multiple schedules in phase I trials were introduced in Günhan et al[7], here we adapt it for sequential investigation of multiple schedules. The time-to-first DLT events are modeled using a time-varying (non-homogeneous) Poisson process. The hazard function is assumed to depend on an exposure measure of the drug (E(t)):

$$h(t) = \beta\, E(t) \tag{1}$$

where β is the only parameter to estimate in the model.

The exposure measure is calculated using a pseudo-PK model which consists of two ordinary

differential equations:

$$\frac{dC(t)}{dt} = -k_e\, C(t) \quad \text{and} \quad C(0) = 0$$
$$\frac{dC_{\text{eff}}(t)}{dt} = k_{\text{eff}}\,(C(t) - C_{\text{eff}}(t)) \quad \text{and} \quad C_{\text{eff}}(0) = 0.$$

where $C(t)$ and $C_{\text{eff}}(t)$ are the concentrations of drug in the central compartment and in the so-called effect compartment, respectively. Due to non-identifiability, the volume in both compartments is set to unity by convention here. Furthermore, k_e is the elimination rate constant and k_{eff} is the PK parameter which governs the delay between the concentration in the central compartment and the concentration in the effect compartment. The parameter k_e is parametrized using the elimination half-life T_e, that is $k_e = \frac{\log(2)}{T_e}$. The parameters k_e and k_{eff} are assumed to be known from previous analyses, for example from another previously studied indication or pre-clinical data.

TITE-PK uses an adapted EWOC criterion. For this purpose, the measure of the interest is the probability of a patient experiencing at least one DLT within the first cycle (shortly the end-of-cycle 1 DLT probability), $P(T \leq t^*|d, f)$, where d and f refer to the dose and frequency of administration, respectively. Using basic event history analysis[18], we have the following equation

$$P(T \leq t^*|d, f) = 1 - e^{-H(t^*|d,f)}, \tag{2}$$

which describes the relationship between the end-of-cycle 1 probabilities and the cumulative hazard function $H(t)$. All patients without a DLT up to the end of cycle 1 will be censored at the end of cycle 1, and patients with a DLT are censored at the time of a DLT. Using Equation (2), it can be shown that

$$\text{cloglog}(P(T \leq t^*|d, f)) = \log(\beta) + \log(\text{AUC}_E(t^*|d, f)) \tag{3}$$

where $\text{cloglog}(x) = \log(-\log(1-x))$ and $\text{AUC}_E(t)$ is the area under the curve of the exposure measure over time.

To help prior specification, $E(t)$ is obtained by scaling $C_{\text{eff}}(t)$ using a reference schedule (reference dose d^* and frequency f^*) at the end of the first treatment cycle (cycle 1: t^*) such that

$$\text{AUC}_E(t^*|d^*, f^*) = 1. \tag{4}$$

By combining Equation (3) and Equation (4), it follows that for the reference schedule $\text{cloglog}(P(T \leq t^*|d^*, f^*)) = \log(\beta)$, which we use for the prior specification of the β parameter. This relationship suggest to constrain β to be positive, which ensures that $h(t) \geq 0$, since $E(t) \geq 0$ for all t (see Equation (1)).

The posterior distributions of end-of-cycle 1 DLT probabilities are classified into three categories in order to inform dose-escalation decisions:

(i) $P(T \leq t^*|d, f) < 0.20$ Underdosing (UD)

(ii) $0.20 \leq P(T \leq t^*|d, f) \leq 0.40$ Targeted toxicity (TT)

(iii) $P(T \leq t^*|d, f) > 0.40$ Overdosing (OD)

The EWOC criterion is fulfilled, if the overdosing probability $P(P(T \leq t^*|d, f) > 0.40)$ is smaller than the feasibility bound a. As the feasibility bound, we use 0.25, which is recommended by Babb

Table 2: Scenarios 1-6 in the simulation study. Doses with dose limiting toxicities in the targeted toxicity interval (0.20 - 0.40) are in boldface. Scenarios 1-6 represent phase I trials with one schedule, that is daily schedule.

	Doses in mg/m^2					
Scenario	2.5	5	7.5	10	12.5	15
1	0.05	0.10	**0.20**	**0.30**	0.50	0.70
2	**0.30**	**0.40**	0.52	0.61	0.76	0.87
3	0.05	0.06	0.08	0.11	0.19	**0.34**
4	0.06	0.08	0.12	0.18	**0.40**	0.71
5	0.10	**0.22**	**0.31**	0.45	0.60	0.72
6	0.50	0.55	0.61	0.69	0.76	0.87

et al[4]. Analogous to the monotonicity of dose-DLT probability assumption of CRM, TITE-PK assumes the monotonicity of the exposure measure and the end-of-cycle 1 DLT probability. That is, $\text{AUC}_E(t^*|d, f)$ is proportional to the end-of-cycle 1 DLT probabilities.

In the case of sequential investigation of multiple schedules, initially TITE-PK is used to conduct the phase I trial with S_1 until the MTD is declared or trial is stopped since all doses are found to be too toxic. In this step, the frequency of administration is the same for dose-escalation decisions. Then, cohorts are recruited with Schedule S_2. For dose-escalation decisions, the information from the phase I trial with S_1 is treated as data together with the new information generated from the phase I trial with S_2. Since TITE-PK is an exposure-response model, there is no need to re-scale the doses from different schedules to make them comparable. As opposed to BLRM MAP and B-CRM methods, data from the completed trials is treated as part of the data instead of as part of the prior distribution.

2.2 Software implementation

We implemented TITE-PK in `Stan`[19] via **rstan** R package, which employs a modern Markov chain Monte Carlo engine. For the application and simulations, four parallel chains of 1,000 MCMC iterations after warm-up of 1,000 iterations are generated. Convergence diagnostics are checked using the Gelman-Rubin statistics and traceplots in the application. There were no divergences reported for the implementation of the application. The R and Stan code to analyze the everolimus application is publicly available from Github (https://github.com/gunhanb/TITEPK_sequential). The main programming code is the Stan code from the linked folder, which conducts the Bayesian computation to calculate posterior distributions. The method can be applied by changing R-code based on the application, for example different doses or schedules, while keeping the Stan code.

3 Results

3.1 Simulation study

We compared the operating characteristics of TITE-PK and alternative methods in a simulation study. The simulation study follows the clinical scenario evaluation framework introduced by Benda et al[20] and it is inspired by the everolimus trial. Firstly, we considered scenarios only involving one schedule to compare the performance of TITE-PK to CRM and BLRM. For the CRM implementation, we used a one-parameter power model via the R package `bcrm`[21]. Both TITE-PK and BLRM recommends the highest dose among the doses which satisfy the EWOC criteria, while CRM recommends the

Table 3: Scenarios 7-13 in the simulation study. Daily doses with dose limiting toxicities in the targeted toxicity interval (0.20 - 0.40) are in boldface.

		Doses with Schedule S_1						**Doses with Schedule S_2**					
Scenario	Schedule	2.5	5	7.5	10	12.5	15	2.5	5	7.5	10	12.5	15
7	S_1	0.05	0.07	0.09	0.10	0.13	0.18						
	S_2							0.08	0.12	0.16	0.18	**0.23**	**0.27**
8	S_1	0.08	0.12	0.16	**0.20**	**0.23**	**0.27**						
	S_2							0.18	**0.26**	**0.34**	0.45	0.49	0.55
9	S_1	0.03	0.12	**0.28**	**0.40**	0.54	0.62						
	S_2							**0.20**	**0.30**	0.45	0.50	0.60	0.75
10	S_1	0.10	**0.20**	**0.34**	**0.40**	0.49	0.55						
	S_2							**0.35**	**0.40**	0.45	0.57	0.67	0.80
11	S_1	0.05	0.07	0.09	0.15	**0.22**	**0.28**						
	S_2							**0.30**	**0.35**	0.48	0.52	0.61	0.70
12	S_1	0.45	0.50	0.55	0.65	0.75	0.85						
	S_2							0.48	0.56	0.62	0.70	0.80	0.88
13	S_1	0.18	**0.26**	**0.34**	0.45	0.49	0.55						
	S_2							0.08	0.12	0.16	0.18	**0.23**	**0.27**

dose which has a DLT probability closest to the target probability. TITE-PK and CRM have one parameter, while BLRM has two parameters in the model. Daily doses of 2.5, 5, 7.5, 10, 12.5, and 15 (mg/m^2) are investigated. The starting dose is 2.5 mg/m^2 for all methods. Scenarios 1-6 are summarized in Table 2. Doses within the targeted toxicity intervals (0.20 - 0.40) are varied based on the scenarios. Scenario 6 is an extreme scenario, where all doses are in the overdosing interval.

We also consider Scenarios 7-13 representing sequential phase I trials with two schedules. In the first step, doses of 2.5, 5, 7.5, 10, 12.5, 15 (mg/m^2) with the dosing frequency of 48 hours (S_1) and in the second step, doses of 2.5, 5, 7.5, 10, 12.5, 15 (mg/m^2) with daily dosing (S_2) are administered. The starting dose for Schedule S_1 is 2.5 mg/m^2. For Schedule S_2, the MTD declared for S_1 is used as the starting dose. Scenarios 7-13 are summarized in Table 3 and displayed in Figure 1. Doses from Schedule S_2 with DLT probabilities within the targeted toxicity intervals (0.20 - 0.40) and discrepancy between DLT probabilities of two schedules are varied based on the scenarios. Scenario 11 is a scenario in which the discrepancy of dose-toxicity curve between the schedules is higher than other scenarios. All doses are in the overdosing interval in Scenario 12. Scenario 13 is Scenario 8 with DLT probabilities for Schedules 1 and 2 switched. Hence, the monotonicity assumption of the exposure and DLT probabilities is violated in Scenario 13. In other words, for the same dose, toxicity is higher with the lower frequent administration. Note that the weekly doses from everolimus are not chosen in the simulations in order to better investigate the monotonicity assumption of DLT probability and exposure. This is because, with the described doses and schedules in the simulations, we can easily vary the order of DLT probability of the same dose with different schedules in the scenarios.

We consider three methods for sequential phase I trial scenarios (Scenarios 7-13): TITE-PK, Bridging CRM (B-CRM), BLRM using MAP prior (BLRM MAP). As explained in the introduction, a sequential phase I trial consists of two steps. In B-CRM, the first step is conducted using the CRM, whereas BLRM is used for the first step of BLRM MAP. In B-CRM, multiple skeletons are constructed using the data from Schedule S_1. The Bayesian model averaging is used to estimate toxicity probabilities with multiple skeletons and to inform the dose-escalation decisions. We used

the publicly available R-code which is provided as the supplementary material of Liu et al[8]. Dose skipping is not allowed in B-CRM. We refer to Liu et al[8] for more details of B-CRM. In BLRM MAP, a meta-analytic-predictive (MAP) prior is created based on the data from Schedule S_1. The MAP prior is used to construct the prior for the parameters of the BLRM. BLRM MAP uses the EWOC criterion to avoid to impose more patients to the overly toxic doses. For BRLM and BLRM MAP, the feasibility bound of 0.25 is used, as recommended in Neuenschwander et al[3]. Full description of BLRM MAP is given in Neuenschwander et al[9]. In TITE-PK and BLRM MAP, dose-escalation by more than 100% mg/m^2 is not allowed. The R package `OncoBayes2`[22] can be used to implement BLRM MAP.

For TITE-PK, we need to determine PK parameters. By mimicking the everolimus trial, PK parameters are chosen as follows. The elimination rate constant is taken as $k_e = \frac{\log(2)}{30}$ (1/h). For k_{eff}, an estimate is derived using the cycle length and the absorption rate. Specifically, a log-normal distribution is constructed by matching the inverse of cycle length 1/504 (1/h) and the absorption rate 2.5 (1/h) as the 0.025 and 0.975 quantiles, respectively. This gives a log-normal distribution with mean parameter 0.37, hence we assume that $\log(k_{\text{eff}}) = 0.37$.

Prior skeletons and distributions are constructed so that prior DLT probabilities from different methods are similar. For TITE-PK model, reference dose and reference dosing frequency are determined using 7.5 mg/m^2 ($d^* = 7.5$ mg/m^2) and 24 hours ($f^* = 1/24$ 1/h). A normal weakly informative prior (WIP) is chosen such that $\log(\beta) \sim \mathcal{N}(\text{cloglog}(P(T \leq t^*|d^*, f^*) = 0.30), 1.25^2)$. This implies that prior median of DLT probability at the reference dose and frequency is 0.30. For BLRM MAP, we choose a WIP assuming median DLT probability of 0.30 at dose 7.5 mg/kg. More specifically, we choose a bivariate normal distribution $(\log(\alpha_1), \log(\alpha_2)) \sim \text{BVN}(\mathbf{m}, \Sigma)$ with means $m_1 = \text{logit}(0.30)$ and $m_2 = 0$, standard deviations $\sigma_1 = 2$ and $\sigma_2 = 1$, and correlation $\rho = 0$. The target probability for the CRM is 0.30, that is the midpoint of the targeted toxicity interval (0.20 - 0.40). For the CRM, the prior skeleton is calculated using the method of Lee and Cheung[23] assuming an indifference interval of 0.10, which produces (0.02, 0.12, 0.30, 0.50, 0.68, 0.80). A normal prior with mean 0 and standard deviation 2 is used as the prior for the power parameter α in the CRM and B-CRM ($\alpha \sim \mathcal{N}(0, 2^2)$), as suggested by Liu et al[8].

The following simulation settings and decision rules are used for TITE-PK, BLRM and BLRM MAP. The maximum number of patients per trial was set to 60. If all doses are in the overdosing interval based on the EWOC criterion, the trial is stopped without selecting any dose as the MTD. Otherwise, the trial continues until the recommendation of the MTD. The recommended MTD must meet the following conditions:

(i) At least 6 patients have been treated at the MTD.

(ii) A minimum of 21 patients have already been treated in the trial.

For the CRM and B-CRM, the trial is terminated for safety, if the following rule is satisfied: $P(\pi_1 > 0.30) < 0.90$ where π_1 is the DLT probability of the lowest dose. The sample size of 21 patients is used unless the trial is stopped due to the safety. For all methods in the simulations, cohort sizes of 3 are used and data for 1,000 trials were generated per scenario.

3.2 Simulation results

The simulation results for Scenarios 1-6 are summarized in Table 4. We calculated six different metrics to evaluate the performance of different methods. Scenarios 1-6 represent phase I trials with one schedule investigated. In Scenario 1, TITE-PK slightly outperforms other methods in terms of recommending the MTD in the targeted toxicity interval. The corresponding percentages are 78% for

Table 4: Simulation results for TITE-PK, CRM, and BLRM in Scenarios 1-6.

	Scenario					
	1	2	3	4	5	6
Probability of selecting MTD in the targeted toxicity interval						
TITE-PK	0.78	0.52	0.75	0.36	0.71	n/a
CRM	0.73	0.61	0.24	0.22	0.79	n/a
BLRM	0.75	0.49	0.64	0.14	0.78	n/a
Probability of selecting MTD in the overdosing interval						
TITE-PK	0.11	0.03	n/a	0.06	0.17	0.11
CRM	0.09	0.04	n/a	0.04	0.10	0.14
BLRM	0.06	0.02	n/a	0.04	0.10	0.07
Probability of selecting no combination as MTD						
TITE-PK	0.01	0.42	0.00	0.01	0.04	0.87
CRM	0.01	0.36	0.01	0.01	0.03	0.86
BLRM	0.01	0.48	0.01	0.01	0.04	0.92
Mean number of patients enrolled						
TITE-PK	24.7	15.4	23.3	27.0	22.8	8.1
CRM	20.9	15.7	20.9	20.8	20.5	8.9
BLRM	23.6	14.9	24.2	24.8	21.9	7.3
Proportion of patients enrolled in the overdosing interval						
TITE-PK	0.28	0.15	n/a	0.13	0.27	1.00
CRM	0.05	0.05	n/a	0.01	0.06	1.00
BLRM	0.10	0.08	n/a	0.11	0.11	1.00
Proportion of DLT observed						
TITE-PK	0.28	0.38	0.21	0.25	0.30	0.52
CRM	0.18	0.33	0.11	0.15	0.22	0.51
BLRM	0.21	0.35	0.15	0.20	0.24	0.50

TITE-PK, 75% for BLRM and 73% for CRM. Also, BLRM yields slightly lower percentage for the MTD selection in the overdosing interval compared to TITE-PK and CRM. BLRM selects the MTD in the overdosing interval in 6% of the time, while TITE-PK and CRM do this in 11% and 9% of the time, respectively. In Scenario 2, CRM yields higher percentage for the MTD selection in the targeted toxicity interval compared to the TITE-PK and BLRM. CRM recommends the MTD in the targeted toxicity interval in 61% of the time, while TITE-PK and BLRM do this in 52% and 49% of the time, respectively. Three methods perform similarly in terms of recommending the MTD in the overdosing interval. In scenario 3, TITE-PK results in the best performance in terms of the MTD selection in the targeted toxicity interval. TITE-PK recommends the MTD in the targeted toxicity interval 75% of the time, while BLRM and CRM do this in 64% and 24% of the time, respectively.

In scenario 4, all methods perform poorly in terms of selecting the MTD in the targeted toxicity, while TITE-PK results in the best performance. TITE-PK yields 36% percentage for the MTD selection in the targeted toxicity interval, while CRM and BLRM yields 22% and 14%, respectively. In scenario 5, CRM (79%) and BLRM (78%) produces slightly higher percentages than TITE-PK (71%) in terms of the selecting MTD in the targeted toxicity interval. In scenario 6, all doses are in the overdosing interval. BLRM (92%) stops the trial with slightly higher percentages compared to CRM (86%) and TITE-PK (87%).

In Scenarios 1, 3, 4 and 5, TITE-PK and BLRM enrolls slightly higher number of patients and results in slightly higher proportions of DLT observed in comparison to CRM. Overall, none of the methods shows superior performance in terms of the investigated metrics. The results depend on the scenarios. Similar results from the comparison of BLRM and CRM was also obtained by the simulation studies in Neuenschwander et al[3].

We continue with Scenarios 7-13 in which sequential phase I trials are investigated. The simulation results under Scenarios 7-13 are summarized in Table 5. In Scenario 7, BLRM MAP produces the best performance in terms of the MTD selection in the targeted toxicity interval, while TITE-PK is the second. The corresponding percentages are 95%, 90%, and 83% for BLRM MAP, TITE-PK, and B-CRM respectively. In Scenarios 8-11, TITE-PK demonstrates superior performance in terms of selecting the MTD in the targeted toxicity interval. TITE-PK selects the MTD in the targeted toxicity interval in 14%, 17%, 16%, and 10% more simulated trials in comparison to the second best performed method in Scenarios 8-11, respectively. In Scenarios 8 and 9, TITE-PK produces lower percentages in terms of the MTD selection in the overdosing interval, selecting MTD in 16% and 3% less simulated trials compared to BLRM MAP. In Scenario 11, CRM (28%) displays superior performance in terms of the MTD selection in the overdosing interval in comparison to other methods. In Scenario 12, TITE-PK and BLRM MAP displays better performance than B-CRM by stopping the trial in 98% and 97% of the time, while requiring less patients than other methods. The monotonicity assumption of the exposure and DLT probabilities is violated in Scenario 13. In Scenario 13, B-CRM outperforms other methods by selecting MTD in the targeted toxicity interval in 22% more trials compared to the BLRM MAP. TITE-PK (17%) displays the worst performance in terms of the MTD selection in the targeted toxicity interval.

In Scenarios 7-13 except 12, different methods enrolls similar number of patients. In Scenarios 7-13 except 12, in terms of the proportion of DLT observed, all methods perform similarly. In Scenarios 7-12, TITE-PK displays the best or the second best performance in terms of the MTD selection in the targeted toxicity and overdosing intervals. However, TITE-PK clearly shows poor performance in Scenario 13, which is expected, as the monotonicity assumption between exposure and DLT probability is violated.

Table 5: Simulation results for TITE-PK, B-CRM, and BLRM-MAP in Scenarios 7-13.

	Scenario						
	7	8	9	10	11	12	13
Probability of selecting MTD in the targeted toxicity interval							
TITE-PK	0.90	0.70	0.94	0.84	0.62	n/a	0.17
B-CRM	0.83	0.50	0.64	0.60	0.52	n/a	0.77
BLRM MAP	0.95	0.56	0.77	0.68	0.46	n/a	0.55
Probability of selecting MTD in the overdosing interval							
TITE-PK	n/a	0.22	0.05	0.02	0.37	0.02	n/a
B-CRM	n/a	0.38	0.08	0.00	0.28	0.25	n/a
BLRM MAP	n/a	0.40	0.21	0.10	0.41	0.03	n/a
Probability of selecting no combination as MTD							
TITE-PK	0.00	0.02	0.01	0.14	0.00	0.98	0.15
B-CRM	0.00	0.02	0.02	0.28	0.20	0.75	0.00
BLRM MAP	0.00	0.02	0.02	0.22	0.12	0.97	0.01
Mean number of patients enrolled							
TITE-PK	21.7	21.7	21.4	19.4	21.8	3.7	19.7
B-CRM	21.0	21.0	21.0	18.0	19.0	9.0	21.1
BLRM MAP	21.5	23.6	21.6	20.0	22.8	4.8	23.4
Proportion of patients enrolled in the overdosing interval							
TITE-PK	n/a	0.39	0.17	0.12	0.61	1.00	n/a
B-CRM	n/a	0.46	0.15	0.06	0.72	1.00	n/a
BLRM MAP	n/a	0.59	0.40	0.26	0.70	1.00	n/a
Mean number of of DLT observed							
TITE-PK	5.3	8.2	6.2	7.5	10.2	1.8	2.4
B-CRM	4.5	7.0	7.3	7.5	8.5	4.0	3.0
BLRM MAP	5.7	9.7	7.7	8.2	11.1	2.4	3.9

3.3 Revisiting the everolimus trial

Returning to the data set described before, consider the everolimus trial shown in Table 1. Firstly, we analyse the data only from the daily schedule using the BLRM, the CRM, and the TITE-PK. Secondly, we analyse it as if the trial is conducted sequentially, specifically S_1 is weekly schedule and S_2 is daily schedule using BLRM MAP, B-CRM, and TITE-PK. The reference schedule is determined using dosing amount of 5 mg/m^2 ($d^* = 5$ mg/m^2) and dosing frequency of 24 hours ($f^* = 1/24$ 1/h). For TITE-PK, PK parameters are chosen such that $T_e = 30$ (hours) and $\log(k_{\text{eff}}) = 0.37$ as explained in the simulation study.

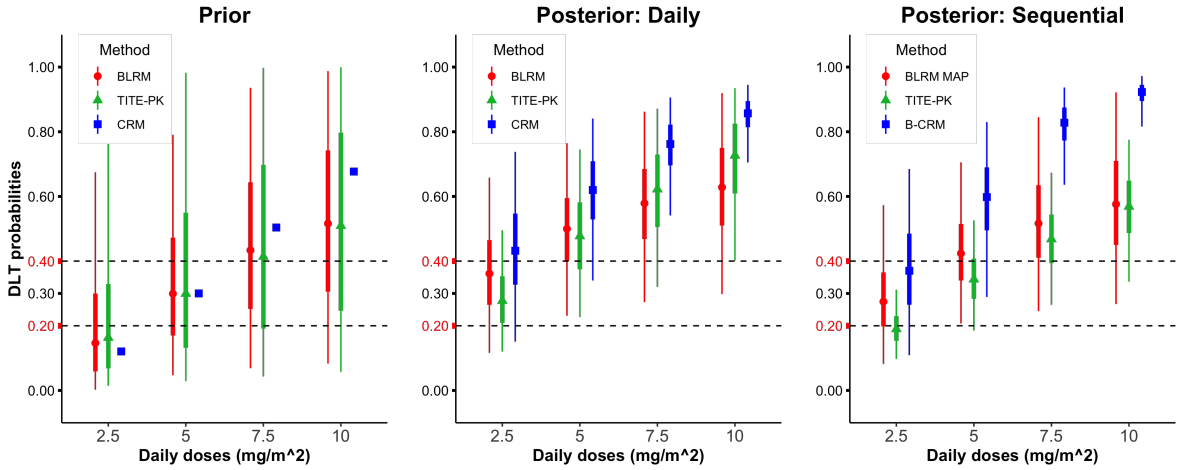

Figure 1: Everolimus trial: Prior medians (A), posterior medians daily (B), and sequential (C), 50% equi-tailed credible intervals (thick lines), and 95% equi-tailed credible intervals (thin lines) of daily doses for DLT probabilities obtained by BLRM (BLRM-MAP for Sequential), CRM (B-CRM for Sequential), and for end-of-cycle 1 DLT probabilities obtained by TITE-PK. Prior skeletons are shown for CRM in the plot A. "Sequential" refers that analysis is done by assuming the trial is conducted sequentially, namely firstly weekly schedule, secondly daily schedule. Also, "Daily" means data only from daily schedule is considered. Vertical dashed lines (0.20-0.40) are the boundaries of the targeted toxicity interval.

To compare BLRM, CRM and TITE-PK models, priors are constructed so that prior DLT probabilities are similar. To define a WIP for BLRM, we choose a bivariate normal prior with following parameters $(m_1 = \text{logit}(\pi_{d^*} = 0.30), m_2 = 0, \sigma_1 = 1.25, \sigma_2 = 1, \rho = 0)$. For the CRM, we use the target probability of 0.30. The prior skeleton is, then, calculated assuming an indifference interval of 0.10, which produces (0.12, 0.30, 0.50, 0.68). For TITE-PK, a normal WIP is chosen such that $\log(\beta) \sim \mathcal{N}(\text{cloglog}(P(T \leq t^*|d^*, f^*) = 0.30), 1.25^2)$. The summaries of prior DLT probabilities of BLRM and TITE-PK, and prior skeletons of CRM are shown in Figure 2A. Points, thick lines and thin lines correspond to median estimates, the 50% and the 95% equi-tailed credible intervals, respectively. Vertical dashed lines (0.20-0.40) are the boundaries of the targeted toxicity interval. Recall that, in TITE-PK and BLRM, eligible doses are determined based on the EWOC criterion, whereas CRM selects the dose closest to the target probability.

Figure 2B displays the posterior estimates of DLT probabilities, when we only consider daily schedule data. BLRM suggests that all doses are in the overdosing interval, meaning that the trial should be stopped without any dose declared as the MTD. The estimated overdosing probability of 2.5 mg/m^2 is 0.40, which is higher than 0.25. For TITE-PK, only 2.5 mg/m^2 is not in the overdosing interval. The overdosing probability of 2.5 mg/m^2 is 0.14, $P(P(T \leq t^*|d = 2.5, f = 24) > 0.40) =$

0.14, which is smaller than 0.25. Although median DLT probability estimate of CRM is higher than the median DLT probability estimate of BLRM, CRM does not conclude that the trial should be stopped. This is because, $P(\pi_1 > 0.30) = 0.80$, which is smaller than 0.90. Furthermore, credible intervals obtained by the CRM is getting shorter with the increasing dose, which was also observed by Neuenschwander et al[3]. Overall, high overdosing probabilities for all doses seem reasonable, since 2 DLT were observed in the 4 patients with 2.5 mg/m^2, and 3 DLT were in the 6 patients with 5 mg/m^2 dose.

We continue by treating the data from the weekly schedule as the completed trial in a sequential phase I trial. We estimate the DLT probabilities of daily doses, but also taking into consideration the data coming from the weekly data. To implement BLRM-MAP, the MAP prior is calculated based on the weekly data. Later, the BLRM is fitted and posterior estimates of DLT probabilities are obtained. In the B-CRM, prior skeletons are calculated using the weekly data. Then, CRM via a Bayesian model averaging method is used to estimate DLT probabilities. TITE-PK, naturally, combines information from different schedules. Figure 2C displays the estimated posterior summaries of DLT probabilities of daily doses obtained by TITE-PK, BLRM-MAP and B-CRM approaches. For both TITE-PK and BLRM-MAP, the overdosing probability of dose 2.5 mg/m^2 is decreased substantially, namely from 0.40 to 0.18 for BLRM-MAP, and from 0.14 to 0.00 for TITE-PK. For CRM, the probability $P(\pi_1 > 0.30)$ is also decreased from 0.80 to 0.67. The reduction of the overdosing probabilities of 2.5 mg/m^2 seems reasonable, since in the weekly schedule data, no DLT were observed in the 5 patients with 20 mg/m^2 and 4 DLT were in the 13 patients with 30 mg/m^2. The interval estimates of 2.5 mg/m^2 and 5 mg/m^2 obtained by TITE-PK are shorter, hence more precise estimates compared to BLRM-MAP and B-CRM. All three methods suggest that daily 2.5 mg/m^2 is sufficiently safe, hence it can be declared as the MTD which was the conclusion of the original phase I trial.

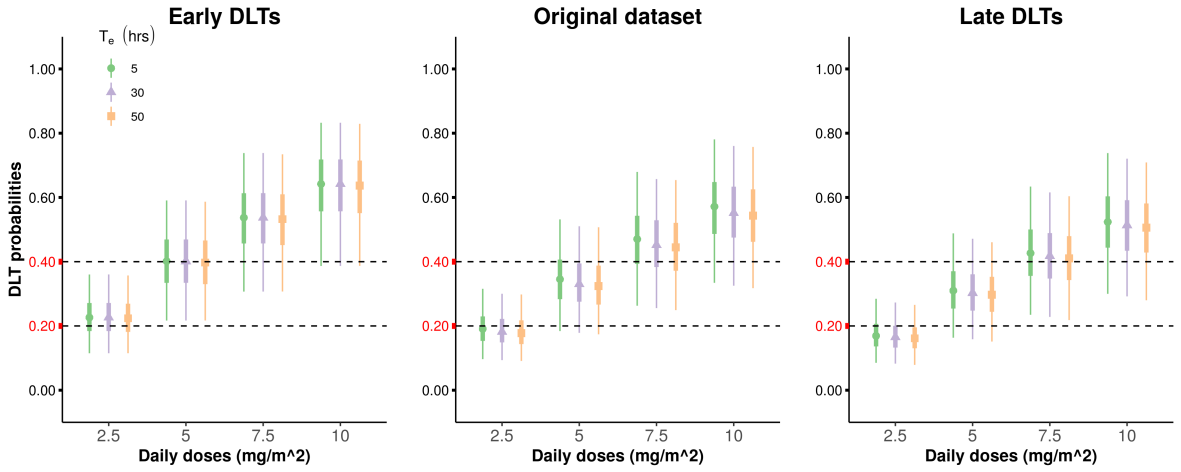

Figure 2: Misspecification of elimination half-life T_e and different timing of DLT. Using different values of T_e, posterior median, 50% and 95% equi-tailed credible intervals for end-of-cycle 1 DLT probabilities obtained by TITE-PK for two hypothetical datasets (early DLT and late DLT) and the original everolimus trial dataset are shown. Early DLT dataset and late DLT dataset are created by changing timing of DLT from day 15 to day 1.5 and to day 20.5, respectively. Data from both weekly and daily schedules are included in the analysis.

As pointed out in Methods Section, by construction of TITE-PK, the elimination half-life T_e is treated as known. To investigate the influence of misspecification of the T_e parameter, we fit TITE-PK using T_e ranging from 5 to 50 hours. The timing of all DLT (in total 9 DLT) were reported at day 15. To examine what would be the influence of the timing of DLT, we also fit TITE-PK to two

hypothetical datasets. Early DLT dataset and late DLT dataset are created by changing timing of DLT from day 15 to day 1.5 and to day 20.5, respectively. Posterior estimates of DLT probabilities for different T_e values and for different timing of DLT are shown in Figure 3. The middle plot corresponds to the original everolimus trial data. Firstly, the posterior medians and credible intervals obtained by different T_e values look very similar. In practice, a reliable estimate of elimination half-life is often not available. Hence, these results are reassuring for the practicality of TITE-PK. Secondly, timing of DLT has a crucial affect on the posterior estimates, and hence the overdosing probabilities. Having the same number of DLT, the earlier the DLT happened, the higher the overdosing probability of the corresponding dose estimated. This makes sense, since one would expect the drug to be more toxic if DLT happened earlier than later.

4 Discussion

In this manuscript, we have adapted TITE-PK for efficiently estimating the maximum tolerable dose in sequential phase I trials involving multiple schedules. To integrate data from different schedules, TITE-PK makes use of exposure-response modelling considering kinetic drug properties. Moreover, we have demonstrated that TITE-PK can be used as an alternative to the standard methods like the BLRM or CRM to conduct phase I trials with only one schedule. In these trials, we have demonstrated that TITE-PK displays similar performance compared to CRM and BLRM. In scenarios with sequential phase I trials, TITE-PK mostly displays superior performance in terms of acceptable dose recommendations in comparison to the bridging CRM and BLRM using MAP approach. An application involving weekly and daily schedules is used to illustrate TITE-PK. Also, using the application, we have shown that TITE-PK is robust against the misspecification of the PK parameter elimination half-life.

Here, we considered a sequential trial in which trial with schedule S_1 is already competed. Another type of a sequential trial can be designed to use the so-called concurrent co-data [9]. That is, the trial with Schedule S_1 is still ongoing, and we would like to utilize the information from the Schedule S_1 to inform dose-escalation decisions with Schedule S_2 (and vice versa). TITE-PK can be used for such designs as well. We did not investigate these situations, since these are beyond the scope of the paper.

In a sequential phase I trial, strata sometimes refer to other than schedules, e.g. patient populations. In such situations, the integration of different strata can be achieved using a MAP approach. Since TITE-PK is parametrized by mimicking the interpretable parameters of the BLRM, it can be extended to use a MAP approach like the BLRM. A key strength of the TITE-PK approach is its ability to integrate the data from different treatment schedules in a model based approach. This makes ad-hoc approaches like dose re-scaling obsolete which reduces the need for strong discounting of historical data from different schedules. However, discounting may still be needed to account for other sources like different patient populations. Recently, Li and Yuan [11] introduced a method to find the MTD for paediatric dose-escalation trial by incorporating information from the concurrent adult data. Their method is based on the CRM and uses Bayesian model averaging to control discounting from the adult data. The BLRM MAP approach makes the assumption of the exchangeability between different schedules. Instead of using a MAP prior, one can use exchangeability/non-exchangeability (EX-NEX) [24,22] approach for phase I trials with multiple schedules, which relaxes the exchangeability assumption.

The monotonicity assumption of the exposure and DLT probabilities is often very reasonable but could be considered a limitation of TITE-PK. Similarly, the BLRM and the CRM assumes the monotonicity of the doses and DLT probabilities. Since, we have used a linear PK model within TITE-PK, the monotonicity of the exposure and DLT probabilities implies the monotonicity of the

dose and DLT probabilities. In the simulations where we investigated phase I trials with one schedules (Scenarios 1-6), we assumed the monotonicity of dose and DLT probabilities. When there is a heavy violation of the assumption of the monotonicity (as in Scenarios 13), the operating characteristics are expected to be weaker compared to bridging CRM or BLRM MAP. The violation of the assumptions occurred, since there is a clear conflict in exposure and DLT profiles between different schedules. Such violations can be informed using the external PK data from the ongoing trial. An extension combining TITE-PK with MAP could be more useful for such situations.

Conflict of interest

S.W. and A.S. own Novartis stakes and are employees of Novartis. T.F. is a consultant to Novartis and has served on data monitoring committees for Novartis. Novartis is the manufacturer of everolimus, an everolimus trial was used to motivate and illustrate the investigations presented here (see Section 1.1 and 3.3).

References

[1] Le Tourneau, C., Lee, J., Siu, L.: Dose escalation methods in phase i cancer clinical trials. J Natl Cancer Inst **101**(10), 708–720 (2009)

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

[3] Neuenschwander, B., Branson, M., Gsponer, T.: Critical aspects of the Bayesian approach to phase I cancer trials. Stat Med **27**(13), 2420–2439 (2008)

[4] Babb, J., Rogatko, A., Zacks, S.: Cancer phase i clinical trials: Efficient dose escalation with overdose control. Stat Med **17**(10), 1103–1120 (1998)

[5] Braun, T., Thall, P., H, N., De Lima, M.: Simultaneously optimizing dose and schedule of a new cytotoxic agent. Clin Trials **4**(2), 113–124 (2007)

[6] Wages, N., O'Quigley, J., Conaway, M.: Phase i design for completely or partially ordered treatment schedules. Stat Med **33**(4), 569–579 (2014)

[7] Günhan, BK and Weber, S and Friede, T. A Bayesian time-to-event pharmacokinetic model for phase I dose-escalation trials with multiple schedules. Stat Med. 2020. 1-15. `http://dx.doi.org/10.1002/sim.8703`.

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

[9] Neuenschwander, B., Roychoudhury, S., Schmidli, H.: On the use of co-data in clinical trials. Stat Biopharm Res **8**(3), 345–354 (2016)

[10] Ollier, A and Morita, S and Ursino, M and Zohar, S. An adaptive power prior for sequential clinical trials - Application to bridging studies. Stat Methods Med Res. 2019. `https://doi.org/10.1177/0962280219886609`.

[11] Li, Y and Yuan, Y. PA-CRM: A continuous reassessment method for pediatric phase I oncology trials with concurrent adult trials. Biometrics. 2020; 1–10. `https://doi.org/10.1111/biom.13217`.

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

[13] National Cancer Institute. Everolimus. `https://www.cancer.gov/about-cancer/treatment/drugs/everolimus`. Updated April, 2018. Accessed September, 2018.

[14] U.S. Food & Drug Administration. FDA approves everolimus for tuberous sclerosis complex-associated partial-onset seizures. `https://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm604351.htm`. Updated April, 2018. Accessed September, 2018.

[15] O'Donnell, A., Faivre, S., Burris III, H., Rea, D., Papadimitrakopoulou, V., Shand, N., Lane, H., Hazell, K., Zoellner, U., Kovarik, J., Brock, C., Jones, S., Raymond, E., Judson, I.: Phase i pharmacokinetic and pharmacodynamic study of the oral mammalian target of rapamycin inhibitor everolimus in patients with advanced solid tumors. J Clin Oncol **26**(10), 1588–1595 (2008)

[16] Besse, B., Heist, R., Papadmitrakopoulou, V., Camidge, D., Beck, J., Schmid, P., Mulatero, C., Miller, N., Dimitrijevic, S., Urva, S., Pylvaenaeinen, I., Petrovic, K., Johnson, B.: A phase ib dose-escalation study of everolimus combined with cisplatin and etoposide as first-line therapy in patients with extensive-stage small-cell lung cancer. Ann Oncol **25**(2), 505–511 (2014)

[17] Cheung, Y., Chappell, R.: Sequential designs for phase i clinical trials with late-onset toxicities. Biometrics **56**(4), 1177–1182 (2000)

[18] Kalbfleisch, J., Prentice, R.F.t.m.: The statistical analysis of failure time data, pp. 31–52. New York, NY: John Wiley & Sons, ??? (2002)

[19] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. J Stat Softw **76**(1), 1–32 (2017)

[20] Benda, N., Branson, M., Maurer, M., Friede, T.: Aspects of modernizing drug development using clinical scenario planning and evaluation. Drug Inf J **44**(3), 299–315 (2010)

[21] Sweeting, M., Wheeler, G.: Bcrm: Bayesian Continual Reassessment Method for Phase I Dose-Escalation Trials. (2019). R package version 0.5.4. `https://CRAN.R-project.org/package=bcrm`

[22] Weber, S., Bean, A., Widmer, L.: OncoBayes2: Bayesian Logistic Regression for Oncology Dose-escalation Trials. (2019). R package version 0.6-5. `https://CRAN.R-project.org/package=OncoBayes2`

[23] Lee, S., Cheung, Y.: Model calibration in the continual reassessment method. Clinical Trials **6**(3), 227–238 (2009)

[24] Neuenschwander, B., Wandel, S., Roychoudhury, S., Bailey, S.: Robust exchangeability designs for early phase clinical trials with multiple strata. Pharm Stat **15**(2), 123–134 (2016)

A.3 Shrinkage estimation for dose-response modeling in phase II trials with multiple schedules

The paper is published online in Statistics in Biopharmaceutical Research. It is available from https://doi.org/10.1080/19466315.2020.1850519. Moreover, preprint version (08.05.2020) is publicly available from https://arxiv.org/abs/2005.04261.

Shrinkage estimation for dose-response modeling in phase II trials with multiple schedules

Burak Kürsad Günhan,[1 2] **Paul Meyvisch**,[3] **Tim Friede**[1]

Recently, phase II trials with multiple schedules (frequency of administrations) have become more popular, for instance in the development of treatments for atopic dermatitis. If the relationship of the dose and response is described by a parametric model, a simplistic approach is to pool doses from different schedules. However, this approach ignores the potential heterogeneity in dose-response curves between schedules. A more reasonable approach is the partial pooling, i.e. certain parameters of the dose-response curves are shared, while others are allowed to vary. Rather than using schedule-specific fixed-effects, we propose a Bayesian hierarchical model with random-effects to model the between-schedule heterogeneity with regard to certain parameters. Schedule-specific dose-response relationships can then be estimated using shrinkage estimation. Considering Emax models, the proposed method displayed desirable performance in terms of the mean absolute error and the coverage probabilities for the dose-response curve compared to the complete pooling. Furthermore, it outperformed the partial pooling with schedule-specific fixed-effects by producing lower mean absolute error and shorter credible intervals. The methods are illustrated using simulations and a phase II trial example in atopic dermatitis. A publicly available R package, `ModStan`, is developed to automate the implementation of the proposed method (https://github.com/gunhanb/ModStan).

Keywords: Shrinkage estimation, multiple schedules, Bayesian inference, phase II trials.

1 Introduction

In phase II of any clinical development program, the investigations of the dose-response relationship of a compound is crucial. Usually, there are two main goals of these investigations: (a) establishing a dose-response signal and (b) estimating the dose-response function (Ruberg, 1995). In addition to the dose, a treatment plan of a phase II trial includes the *schedule* (or dose regimen), that is the frequency of the administration, for instance a weekly or biweekly schedule. Recently, phase II trials with multiple schedules have become more popular, for instance in the development of monoclonal antibodies as treatments for a variety of diseases including hypercholesterolaemia (Giugliano et al., 2012) and atopic dermatitis (Thaçi et al., 2016). Eichenfield and Stein Gold (2017) reviewed many therapies for atopic dermatitis which were in phase II or III of clinical development. Multiple schedules were investigated in phase II trials of almost half of the investigated therapies (Eichenfield and Stein Gold, 2017). However, standard methods for dose-response estimation cannot account for multiple schedules.

Estimating separate dose-response curves for each schedule by a parametric model is a one way to tackle this problem, that is full stratification of the dose-response curves. However, this method ignores the information shared between different schedules. Alternatively, one can completely pool doses from different schedules. The main problem with complete pooling is that it does not take into account the potential heterogeneity between different schedules. A more reasonable approach is the *partial pooling*, that is certain parameters of the dose-response curves are shared, while others are

[1] *Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany*
[2] *Correspondence to: Burak Kürsad Günhan; email:* `burak.gunhan@med.uni-goettingen.de`
[3] *Galapagos NV, Mechelen, Belgium*

allowed to vary. A placebo effect parameter can be, reasonably, assumed shared between schedules, whereas this may not be true for the ED_{50} parameter, the dose at which half of the maximum effect is reached. Feller et al. (2017) proposed a partial pooling approach in which unshared parameters are treated as schedule-specific fixed-effects (Möllenhoff et al., 2019).

We consider trials with (very) few schedules of small to moderate size. Here, borrowing available information is of great interest. Rather than using schedule-specific fixed-effects, we propose a Bayesian hierarchical model with random-effects to model the between-schedule heterogeneity with regard to certain parameters. schedule-specific parameters can be estimated using *shrinkage estimation*. The basic idea of the shrinkage estimation is that stratified parameter estimates can be improved by shrinking towards the population mean. It has been shown that the shrinkage estimation improves the estimation accuracy in comparison to estimates obtained by pooling or stratification (Efron and Morris, 1975). Shrinkage estimation in the context of clinical trials were investigated by Jones et al. (2011) and Freidlin and Korn (2013) among others. A popular application is the estimation of the treatment effect in the presence of subgroups, for example estimating response rate in a phase II trial with multiple patient populations (Neuenschwander et al., 2016). Here, we are interested in parametric dose-response models in the presence of multiple schedules, hence shrinkage estimators of the parameters of a dose-response model, for example the ED_{50} parameter of an Emax model. Shrinkage estimation allows *dynamic borrowing* (Viele et al., 2014), in which the weights for each schedule depend on the data instead of using fixed weights. Dynamic borrowing results in considerable gain in efficiency, while being a robust method against the heterogeneity between schedules. A theoretical justification for shrinkage can be established through the concept of *exchangeability* of the parameters between schedules. This means that finding no systematic reason to distinguish schedule-specific parameters, in other words, they are similar, but not identical (Greenland, 2000). Usually, the assumption of exchangebility indicates the schedule-specific parameters come from a common distribution with an overall mean. For the ED_{50} parameter, we assume the re-scaled and log transformed ED_{50} parameter estimates (using the corresponding frequency of each schedule) are exchangeable.

In this manuscript, we propose a Bayesian hierarchical model which utilizes shrinkage estimation for certain parameters of the dose-response model in order to dynamically borrow strength across schedules in a phase II trial. Another contribution is the introduction of a publicly available R package, `ModStan`. In Section 2, two phase II trials with multiple schedules for the treatment of atopic dermatitis are described. We introduce the proposed method to analyze phase II trials with multiple schedules in Section 3. We also describe partial pooling with assuming schedule-specific fixed-effects for certain parameters, discuss the choice of priors and implementation of the proposed method. We evaluated the long-run properties of different methods in a simulation study in Section 4. One of the illustrative applications is revisited to display the proposed method and compare it to the alternatives in Section 5. We close with some conclusions and outlook.

2 Illustrative applications

Atopic dermatitis, the most common form of eczema, is a chronic inflammatory disease that is characterized by skin rash and itching (Mayo Clinic, 2018). Recently, there is an increasing number of clinical trials investigating novel systemic agents for the treatment of atopic dermatitis (Alexander et al., 2019). We consider phase II trials of two human monoclonal antibodies, dupilumab and MOR106, for the treatment of atopic dermatitis. Designs of two trials are listed in Table 1. For both trials, patients were randomized into six arms including a placebo arm. We consider these two trials, since they were both designed to investigate multiple schedules. The dupilumab trial contains three schedules (weekly, biweekly, and monthly), whereas the MOR106 trial contains two (biweekly and

monthly). The placebo doses were administered with the highest frequencies including a weekly and a biweekly schedule for dupilumab and MOR106, respectively. The primary endpoint of both trials is the percentage change from baseline in Eczema Area and Severity Index (EASI) score at Day 85. The EASI scoring system is used to grade the severity of the signs of eczema. EASI scores take values between 0 and 72 and higher EASI score means higher severity. Dupilumab and MOR106 trials are used to motivate our simulation studies in Section 4. The dupilumab trial was completed in September 2014. Multiple comparisons procedure was used as the primary statistical analysis in the dupilumab trial (Thaçi et al., 2016). For the purpose of illustration, we will analyze the dupilumab trial using different modeling approaches in Section 5. In October 2019, the MOR106 trial was terminated due to lack of efficacy in the interim analysis (MorphoSys AG, 2019).

Table 1: Designs of two phase II trials in atopic dermatitis (Dupilumab and MOR106) involving different schedules. Clinicaltrial.gov identifiers are displayed for two trials.

	Dupilumab: NCT01859988			**MOR106**: NCT03568071		
Arm	Schedule	Dose (mg/m^2)	Planned sample size	Schedule	Dose (mg/kg)	Planned sample size
1	Weekly	0	40	Biweekly	0	45
2	Weekly	300	40	Biweekly	1	45
3	Biweekly	200	40	Biweekly	3	45
4	Biweekly	300	40	Biweekly	10	45
5	Monthly	100	40	Monthly	1	30
6	Monthly	300	40	Monthly	3	30

3 Statistical methods

Assume that a response y_{ijk} (an efficacy or a safety outcome) is observed for schedule i, dose j and patient k. Following Feller et al. (2017), we assume a normal likelihood for a continuous outcome:

$$y_{ijk} \sim \mathcal{N}(f(d_j^{(i)}, \boldsymbol{\theta}), \sigma_i^2) \tag{1}$$

where $\boldsymbol{\theta}$ refers to the model parameters and σ_i to the error standard deviation. The $f(d_j^{(i)}, \boldsymbol{\theta})$ represents the functional form of the dose-response relationship for schedule i. Other outcome types, for instance dichotomous or count, can be modeled by specifying appropriate likelihood (e. g. Binomial or Poisson) and the link function (e. g. logit or log transformation).

There are a number of candidate models for the functional form including the popular Emax model (Thomas et al., 2014), that is

$$f(d_j^{(i)}, \boldsymbol{\theta}) = \mathrm{E}_0^{(i)} + \mathrm{E}_{\max}^{(i)} \frac{d_j^{(i)}}{\mathrm{ED}_{50}^{(i)} + d_j^{(i)}} \tag{2}$$

where $\mathrm{E}_0^{(i)}$ is the placebo response and $\mathrm{E}_{\max}^{(i)}$ is the maximum effect attributable to the drug. The $\mathrm{ED}_{50}^{(i)}$ parameter represents the dose at which half of the maximum effect is reached. In the manuscript, we exclusively use the Emax model, see Bretz et al. (2005) for different candidate models.

As explained in the introduction, one way of modeling the dose-response curves is to treat all model parameters as schedule-specific fixed-effects. However, such an analysis is not the most efficient,

when certain aspects of the dose-response curves in different schedules are similar. Alternatively, one can consider a complete pooled analysis in which all model parameters from different schedules are assumed to be the same. This approach is also problematic, since it ignores the potential heterogeneity between dose-response curves of different schedules. A more reasonable approach is the partial pooling (Feller et al., 2017), which strikes a balance between efficiency and robustness. It is often reasonable to assume that placebo effect $\mathrm{E}_0^{(i)}$ is the same for different schedules, that is, $\mathrm{E}_0^{(1)} = \mathrm{E}_0^{(2)} = \ldots$. This is especially the case, when there is only one placebo arm investigated in the trial as in the illustrative trials described in Section 2. In some situations, it might also make sense to assume that the maximum efficacy for high doses is same, $\mathrm{E}_{\max}^{(1)} = \mathrm{E}_{\max}^{(2)} = \ldots$. However, it might not be reasonable to assume the dose providing half of the maximum efficacy is the same for different schedules, that is $\mathrm{ED}_{50}^{(1)} \neq \mathrm{ED}_{50}^{(2)} \neq \ldots$. Feller et al. (2017) suggested to treat the unshared parameters, for example $\mathrm{E}_{\max}^{(i)}$ and/or $\mathrm{ED}_{50}^{(i)}$, as schedule-specific fixed-effects in the partial pooling approach.

3.1 Proposed method: Partial pooling with random-effects

Rather than using schedule-specific fixed-effects, we propose a Bayesian hierarchical model with random-effects to model the between-schedule heterogeneity with regard to certain parameters in the partial pooling approach. In other words, we suggest partial pooling with assuming schedule-specific random-effects for certain parameters of the dose-response model. To be concrete, assume that we want to obtain schedule-specific random-effects for $\mathrm{ED}_{50}^{(i)}$. Firstly, we need to re-scale $\mathrm{ED}_{50}^{(i)}$ parameters. For this reason, we specify a *reference schedule* (i_{ref}). The re-scaled parameters are given by $\mathrm{ED}_{50}^{*(i)} = \mathrm{ED}_{50}^{(i)} \left(f^{(i)}/f^{(i_{\mathrm{ref}})}\right)$ where $f^{(i_{\mathrm{ref}})}$ and $f^{(i)}$ are the frequency of administration of the reference schedule i_{ref} and the schedule i, respectively. The $\mathrm{ED}_{50}^{(i)}$ is modeled on the log-scale, since it is necessarily positive as a dose. We assume that the re-scaled schedule-specific $\mathrm{ED}_{50}^{*(i)}$ estimates are exchangeable

$$\log(\mathrm{ED}_{50}^{*(i)}) \sim \mathcal{N}(\mu_{\mathrm{ED}_{50}}, \tau^2_{\mathrm{ED}_{50}}) \tag{3}$$

where $\mu_{\mathrm{ED}_{50}}$ is the overall mean and $\tau_{\mathrm{ED}_{50}}$ is the between-schedule heterogeneity in $\log(\mathrm{ED}_{50}^{*(i)})$. Our main interest is in the schedule-specific estimates, $\mathrm{ED}_{50}^{(i)}$. If the heterogeneity $\tau_{\mathrm{ED}_{50}}$ is zero, then the model reduces to a model assuming shared $\mathrm{ED}_{50}^{*(i)}$ parameters. Note that the results are invariant to the choice of the reference schedule. Furthermore, similar to the ED_{50} parameter, shrinkage estimates of $\mathrm{E}_{\max}$ parameter can be obtained. There is no need to use the re-scaling or the log transformation for the $\mathrm{E}_{\max}$ parameter. Treating $\mathrm{E}_{\max}$ and/or ED_{50} parameters differently, assuming either one or both of them shared between schedules or assuming schedule-specific random-effects, results in a variety of alternative models.

Complete pooling and partial pooling approaches can be fitted using likelihood estimation. For example, Möllenhoff et al. (2019) demonstrated the likelihood implementation of the partial pooling with assuming schedule-specific fixed-effects for $\mathrm{ED}_{50}^{(i)}$ using constrained nonlinear optimization via `alabama` (Varadhan, 2015) R package. Alternatively, Bayesian approaches can be used, which we consider in this paper.

3.2 Prior distributions

For the Bayesian implementation, we need to specify prior distributions for the model parameters E_0, $\mathrm{E}_{\max}$ $\mu_{\mathrm{ED}_{50}}$, $\tau_{\mathrm{ED}_{50}}$ and σ for the partial pooling assuming schedule-specific random-effects for $\mathrm{ED}_{50}^{(i)}$. We use vague (non-informative) priors, $\mathcal{N}(0, 100^2)$, for the parameters E_0 and $\mathrm{E}_{\max}$, and a half-normal prior with scale 100 for σ, $\mathcal{HN}(100)$. The parameters $\mu_{\mathrm{ED}_{50}}$ and $\tau_{\mathrm{ED}_{50}}$ need special

attention, since the priors of both parameters have strong influence on the posterior estimates. The difficulty of the estimation of the $\tau_{\text{ED}_{50}}$ stems from the small number of the schedules. For example, in our two illustrative trials, there are only two and three schedules available. This is similar to the meta-analysis of few studies, in which the estimation of the between-trial heterogeneity has gained considerable attention in the literature (Gelman, 2006). Friede et al. (2017) suggested the use of *weakly informative priors* (WIP) for the heterogeneity parameter in the case of meta-analysis of few studies, specifically half-normal priors with the scale of 0.5 or 1, when relative measures such as odds ratios, relative risks or hazard ratios (on the logarithmic scale) are used to describe the effect. Inspired by these, we can also construct a WIP for the $\tau_{\text{ED}_{50}}$ to represent plausible range of $\text{ED}_{50}^{*(i)}$ values (Spiegelhalter et al., 2004). The 95% of values of $\log(\text{ED}_{50}^{*(i)})$ will lie in the interval $\mu_{\text{ED}_{50}} \pm 1.96 \cdot \tau_{\text{ED}_{50}}$, hence the 97.5% and 2.5% values of $\log(\text{ED}_{50}^{*(i)})$ are $2 \cdot 1.96 \cdot \tau_{\text{ED}_{50}}$ apart. Accordingly, the ratio of the 97.5% to the 2.5% point of the distribution of $\text{ED}_{50}^{*(i)}$ values is $\exp(3.92 \cdot \tau_{\text{ED}_{50}})$. Table 2 lists the "range" of $\text{ED}_{50}^{*(i)}$ values based on different $\tau_{\text{ED}_{50}}$. In order to cover typical $\tau_{\text{ED}_{50}}$ values conservatively, we will use half-normal priors with scale 1, i.e. $\mathcal{HN}(1)$. When we are interested in the shrinkage estimates of $\text{E}_{\max}$, the construction of the WIP for τ_{Emax} is slightly different. This is because $\text{E}_{\max}$ is computed on the original scale, not on the logarithmic scale. Here, the difference (instead of the ratio) between the 97.5% and the 2.5% point of the distribution of $\text{E}_{\max}$ values is $3.92 \cdot \tau_{\text{Emax}}$. To cover plausible τ_{Emax} values, we will use half-normal priors with the scale 10, $\mathcal{HN}(10)$.

Table 2: Between-schedule heterogeneity $\tau_{\text{ED}_{50}}$ in $\log(\text{ED}_{50}^{*(i)})$: $\tau_{\text{ED}_{50}}$ referring small to very large heterogeneity. The "range", $\exp(3.92 \cdot \tau_{\text{ED}_{50}})$, refers to the ratio of the 97.5% to the 2.5% point of the distribution of $\text{ED}_{50}^{*(i)}$.

$\tau_{\text{ED}_{50}}$	"range" of $\text{ED}_{50}^{*(i)}$
0.125 (small)	1.63
0.25 (moderate)	2.66
0.5 (substantial)	7.10
1 (large)	50.40
2 (very large)	2540.20

The parameter ED_{50} is different from E_0 and $\text{E}_{\max}$ in the sense that it is the only parameter that enters the model non-linearly. In the frequentist framework, it is a common practice to impose bounds (lower and upper bounds) on the space for ED_{50}, since the maximum likelihood estimator (MLE) often does not converge (Bornkamp, 2014). However, the estimate will often exactly equal to the specified upper bound, which is unacceptable. In a Bayesian framework, simple prior choice for the ED_{50} are uniform distributions with finite bounds. However, uniform prior distributions on ED_{50} are problematic, since they strongly depend on the parametrization: One may end up with completely different implied prior distributions for the dose-response curve. A better prior for ED_{50} is the Jeffreys prior, which is invariant to parametrization. It is defined as $p(\boldsymbol{\theta}) \propto \sqrt{|I(\boldsymbol{d}, \boldsymbol{w}, \boldsymbol{\theta})|}$ where $\sqrt{|I(\boldsymbol{d}, \boldsymbol{w}, \boldsymbol{\theta})|}$ is the Fisher information, and $\boldsymbol{w}$ is the vector of proportion of patients allocated at dose $\boldsymbol{d}$. Hence, Jeffreys prior depends on the observed design $(\boldsymbol{x}, \boldsymbol{w})$. One cannot state the Jeffreys prior before data collection, which is crucial in many applications, e.g. in the presence of missing data or two stage designs.

Bornkamp (2012) introduced the *functional uniform prior* which is a modified version of the Jeffreys prior. Functional uniform priors are uniformly distributed on the potential different shapes of the underlying nonlinear model function. These priors are also invariant with respect to parametrization

of the model function and typically result in rather non-uniform prior distributions on the parameter scale. Instead of the actual observed design, functional uniform priors are calculated using a grid of doses as $\boldsymbol{x}$ and equal weights for $\boldsymbol{w}$. More specifically, say, the gradient function of the Emax model is given by $J_x(\boldsymbol{\theta}) = (1, x/(x + \mathrm{ED}_{50}), -x/\mathrm{E}_{\max}/(x + \mathrm{ED}_{50})^2)$. Let $\boldsymbol{x}$ be a grid of doses and $F(\boldsymbol{\theta})$ be the matrix with $J_x(\boldsymbol{\theta})$, x in the rows. Then, the functional uniform prior is proportional to $\sqrt{|Z^*(\boldsymbol{\theta})|}$ where $Z^*(\boldsymbol{\theta}) = F^T(\boldsymbol{\theta})\, F(\boldsymbol{\theta})$ (see Bornkamp (2014) for more detailed explanations). An approximation of the functional uniform prior for ED_{50} is given as the log-normal distribution with mean -2.5 and standard deviation 1.8, when the ED_{50} is re-scaled with the maximum available dose D, that is ED_{50}/D (Bornkamp, 2014). For the simulations and the application, we used the approximation of the functional uniform prior, since it is computationally cheaper. In all models, we use the bounds [0, $1.5 \cdot D$] for the space of ED_{50} (or $\mu_{\mathrm{ED}_{50}}$) parameter.

3.3 Implementation of the proposed method

In a Bayesian framework, we fitted the described statistical models using the probabilistic programming language **Stan** which employs a modern Markov chain Monte Carlo (MCMC) algorithm, namely, Hamiltonian Monte Carlo with the No-U-Turn Sampler (Carpenter et al., 2017). The parametrization used for the statistical model influences the MCMC performance. A centered parametrization such as Equation (3) may cause some computational difficulties such as difficulty in convergence in the presence of data sparsity such as meta-analysis of few studies (Betancourt and Girolami, 2015) or dose-response modeling of phase II trials with few schedules. An alternative parametrization, that is a non-centered parametrization, overcomes these computational difficulties. To be more precise, by the reparametrization of the location and scale parameters, Equation (3) becomes $\log(\mathrm{ED}_{50}^{*(i)}) = \mu_{\mathrm{ED}_{50}} + u_i \cdot \tau_{\mathrm{ED}_{50}}$ where $u_i \sim (0, 1)$ (Günhan et al., 2020). The **Stan** code defining the partial pooling with schedule-specific random-effects for $\mathrm{ED}_{50}^{(i)}$ is shown in Listing 1.

To facilitate the implementation of our proposed method for the practitioners, we have developed an R package, `ModStan` (https://github.com/gunhanb/ModStan). `ModStan` is a purpose-build package defined on top of the `rstan`, the R interface for **Stan**. We show how to install and use `ModStan` in Appendix A.

4 Simulation study

In order to evaluate the long-run properties of the proposed method and compare it with some alternative methods, a simulation study was conducted.

4.1 Simulation settings and implementation

The scenarios considered are motivated by the dupilumab and MOR106 trials described in Section 2. Each generated trial consists of seven arms: one placebo arm and 1, 3, and 10 mg/kg for both biweekly and monthly schedules. The primary outcome is the percentage change from baseline in EASI score. Hence, the datasets are generated under the assumption of normally distributed outcomes, specifically Equation (1). The underlying dose-response function is assumed to be an Emax model, that is Equation (2). True values for $\mathrm{E}_0^{(i)}$, $\mathrm{E}_{\max}^{(i)}$ and σ_i are taken as -20%, -60%, and 35% for both schedules, respectively. Furthermore, $\mathrm{ED}_{50}^{\text{biweekly}}$ is assumed to be 2 mg/kg. A total of 27 scenarios are obtained by varying the $\mathrm{ED}_{50}^{\text{monthly}}$ ($\mathrm{ED}_{50}^{\text{monthly}} \in \{1, 2, 3, 3.5, 4, 4.5, 5, 6, 10$ (mg/kg)$\}$) and sample sizes of each arm ($N \in \{30, 45, 60\}$). $\mathrm{ED}_{50}^{\text{monthly}}$ values are chosen to investigate the influence of the difference between true values of $\mathrm{ED}_{50}^{\text{biweekly}}$ and $\mathrm{ED}_{50}^{\text{monthly}}$ on the performance. Figure 1 displays different dose-response curves for the monthly schedule investigated in the simulations. When $\mathrm{ED}_{50}^{\text{monthly}}$ corresponds

```
1
2  data {
3    int<lower=1> N_obs;                          // num of observations
4    int<lower=1> N_schedule;                     // num of schedules
5    int<lower=1> N_pred;                         // num of predicted doses
6    real resp[N_obs];                            // responses
7    real<lower=0> dose[N_obs];                   // doses
8    int schedule[N_obs];                         // schedule indicator
9    real<lower=0> freq[N_obs];                   // frequency of administration (hrs)
10 }
11 parameters {
12   real E0;                                     // placebo effect (shared)
13   real Emax;                                   // Emax parameter (shared)
14   real log_ED50_raw[N_schedule];               // re-scaled log(ED50) parameters
15   real<lower=0> sigma;                         // standard deviation for errors
16   real<lower=0, upper=1.5> mu_ED50_raw;        // mean of log(ED50) random-effects
17   real<lower=0> tau_ED50;                      // between-schedule heterogeneity
18 }
19 transformed parameters{
20   real mu_ED50;
21   real log_ED50[N_schedule];
22   real<lower=0> ED50[N_schedule];
23   vector[N_obs] resp_hat;
24
25   mu_ED50 = log(mu_ED50_raw * max(dose));
26   for(i in 1:N_schedule)
27     log_ED50[i] = mu_ED50 + log_ED50_raw[i] * tau_ED50;
28   // Taking exponentials and rescaling ED50 parameters
29   for(i in 1:N_schedule)
30     ED50[i] = exp(log_ED50[i]) * (freq[i]/ freq_ref);
31
32   // Dose-response: Emax model
33   for(i in 1:N_obs)
34     resp_hat[i] = E0 + (Emax * dose[i]) / (ED50[schedule[i]] + dose[i]);
35 }
36 model {
37   // random-effects
38   log_ED50_raw ~ normal(0, 1);  // implies log(ED50) ~ normal(mu_ED50, tau_ED50)
39   // likelihood
40   resp ~ normal(resp_hat, sigma);
41   // prior distributions
42   sigma  ~ normal(0, 100);
43   E0 ~ normal(0, 100);
44   Emax ~ normal(0, 100);
45   // approximation to the functional uniform prior
46   mu_ED50_raw  ~ lognormal(-2.5, 1.8);
47   tau_ED50  ~ normal(0, 1);
48 }
```

Listing 1: **Stan** code defining the partial pooling with schedule-specific random-effects for ED_{50} parameter. The parameters E_0, E_{max} and σ are assumed to be shared between schedules.

to 4 mg/kg, there is no heterogeneity in ED_{50} parameters between schedules. This is because if we re-scale ED_{50}^{monthly} to transform on the biweekly scale (simply dividing by two), we obtain 2 mg/kg, which is the true value of $ED_{50}^{\text{biweekly}}$. Accordingly, when the true value of ED_{50}^{monthly} deviates from 4 mg/kg, the heterogeneity between schedules in ED_{50} increases. The simulations were carried out with 1 000 replications per scenario.

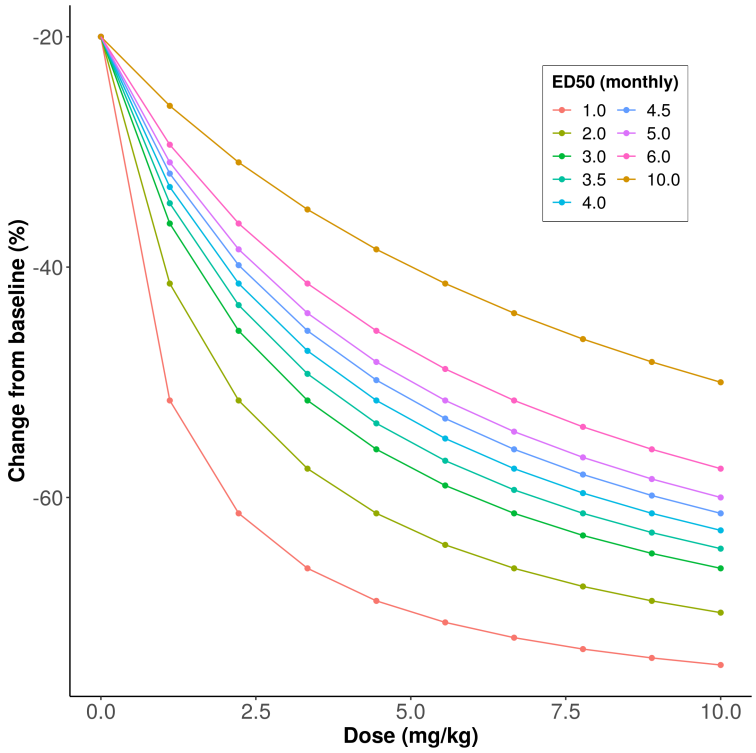

Figure 1: Dose-response curves for the monthly schedule investigated in the simulation study. Different curves are generated by varying ED_{50}^{monthly} parameter value.

In the proposed method, we assume that $E_0^{(i)}$, $E_{\max}^{(i)}$ and σ_i are shared between schedules, while $ED_{50}^{(i)}$ are assumed to be schedule-specific random-effects. In other words, the proposed method corresponds to the partial pooling with assuming schedule-specific random-effects for $ED_{50}^{(i)}$ ("PP - RE"). As a comparator, we use the model in which $ED_{50}^{(i)}$ are assumed to be schedule-specific fixed-effects, while other parameters are shared ("PP - FE"). Both partial pooling approaches (PP - RE and PP - FE) are fitted via a Bayesian approach. We also consider the complete pooling method via a frequentist and a Bayesian approach ("CP (Frequentist)" and "CP (Bayesian)"). For the partial pooling with schedule-specific random-effects, we used the biweekly schedule as the reference schedule to re-scale the $ED_{50}^{(i)}$ parameters. To implement the complete pooling approaches, all doses should be re-scaled into the same schedule. For this purpose, we transform the doses from the monthly schedule into the biweekly schedule. Accordingly, the new set of doses becomes $\{0, 0.5, 1, 1.5, 3, 10$ (mg/kg)$\}$ for complete pooling approaches.

The complete pooling (Frequentist) is fitted using `fitMod` function from the `DoseFinding` (Bornkamp et al., 2018) R package. All Bayesian methods are fitted using **Stan** and the prior distributions from Section 3.2 are used. Three MCMC chains were run in parallel for a total of 4 000 iterations including 2 000 iterations of burn-in. Convergence diagnostics are evaluated in some replications, these MCMC settings are chosen accordingly. The ED_{50} parameter is assumed to be within the bounds of $[0.001, 1.5 \cdot 10]$ to ensure identifiability for all methods.

4.2 Simulation results

For each simulation run, we calculated point estimates ($\hat{f}$) for the dose-response function (f) (the pointwise posterior median or the maximum likelihood estimate) at some pre-specified dose levels. For this purpose, ten dose levels are chosen between 0 and 10 mg/kg equidistantly, namely $\text{dose}_l \in \{0.00, 1.11, \ldots, 10.00\}$. Additionally, interval estimates (95% confidence interval or 95% equi-tailed credible intervals) are derived at each dose_l. These computations are done for the dose-response function of the biweekly schedule. The following three performance measures are calculated:

- MAE: Mean absolute error for the dose-response function, $1/10\,\Sigma_{\text{dose}_l=0}^{10}|f(\text{dose}_l) - \hat{f}(\text{dose}_l)|$ at each dose_l.
- Coverage probability: Mean coverage probability of the interval estimates evaluated at each dose_l.
- Mean length: Mean length of the interval estimates at each dose_l.

The lower MAE for the point estimates, the shorter interval estimates, and the coverage probability of 95% for the interval estimates are desirable. The MAE obtained by the four methods is displayed in the first row of Figure 2. Different columns of Figure 2 correspond to different sample sizes N which are investigated in the simulations. Across different sample sizes, the relative performances of the four methods remain similar. The scenario of $\text{ED}_{50}^{\text{monthly}} = 4$ corresponds to the scenario without heterogeneity in the re-scaled $\text{ED}_{50}^{(i)}$ between biweekly and monthly schedules, which is shown by a vertical dashed line. The heterogeneity increases, when $\text{ED}_{50}^{\text{monthly}}$ deviates from 4. Both complete pooling approaches display better performance than both partial pooling approaches in terms of the MAE, when the $\text{ED}_{50}^{\text{monthly}}$ is 4. However, the partial pooling approaches result in more robust performance across $\text{ED}_{50}^{\text{monthly}}$ values in comparison to the pooling approaches. The partial pooling with random-effects uses the prior $\mathcal{HN}(1)$ for the heterogeneity parameter $\tau_{\text{ED}_{50}}$. If we increase the value of the prior standard deviation (that is 1), then the performance of the partial pooling with random-effects will get closer to the partial pooling with fixed-effects. Similarly, if we assume that $\tau_{\text{ED}_{50}}$ equals to zero, the partial pooling with random-effects reduces to, effectively, the complete pooling (Bayesian). The partial pooling with random-effects yields better performance than the partial pooling with fixed-effects in terms of the MAE across different $\text{ED}_{50}^{\text{monthly}}$ values and sample sizes except the most extreme scenarios, namely $\text{ED}_{50}^{\text{monthly}} = 1$ or 10. Note that the main difference between the complete pooling (Bayesian) and complete pooling (Frequentist) is that in the former, functional uniform priors used for $\text{ED}_{50}^{(i)}$ parameters. The small discrepancy between the MAE obtained by the complete pooling (Bayesian) and the complete pooling (Frequentist) can be explained by this difference. Furthermore, when the sample sizes increase, the MAE decreases in the four methods as expected.

Figure 2 also shows coverage probabilities of the interval estimates obtained by the four methods. The complete pooling approaches result in a concave shape and display unacceptably low coverage when $\text{ED}_{50}^{\text{monthly}}$ deviates from 4. This undesirable performance of the complete pooling approaches is more pronounced, when the sample size increases. As in the MAE, both partial pooling approaches show more robust performance in terms of the coverage probabilities in comparison to the complete pooling approaches. The partial pooling with random-effects yields superior performance in terms of the coverage probability compared to the partial pooling with fixed-effects across different $\text{ED}_{50}^{\text{monthly}}$ values and sample sizes except when $\text{ED}_{50}^{\text{monthly}} = 1$. Figure 3 illustrates the ratios of lengths of credible intervals for the dose-response functions obtained by the partial pooling approaches. The denominator of the ratio is the length of the credible interval obtained by the partial pooling with random-effects.

The partial pooling with random-effects results in slightly shorter credible intervals, while it produces slightly higher coverage probability compared to the partial pooling with fixed-effects in most of the scenarios.

To examine the influence of the potential heterogeneity in $E_{max}^{(i)}$ between schedules, we conducted additional simulations. True values for $E_0^{(i)}$ and σ_i are taken as -20% and 35% for both schedules, respectively. The $ED_{50}^{biweekly}$ and $ED_{50}^{monthly}$ are assumed to be 2 and 4 mg/kg, respectively. This corresponds to assuming no heterogeneity in $ED_{50}^{(i)}$ parameters between schedules, since we focused on $E_{max}^{(i)}$ in these simulations. The $E_{max}^{biweekly}$ is assumed to be -60%. Sample size for each arm is 45. Three scenarios are generated by varying $E_{max}^{monthly}$ values ($E_{max}^{monthly} \in \{-70\%, -60\%, -50\%\}$). Notice that when $E_{max}^{monthly} = -60\%$, there is no heterogeneity in $E_{max}^{(i)}$ values. In the partial pooling with fixed-effects, both $E_{max}^{(i)}$ and $ED_{50}^{(i)}$ parameters are treated as schedule-specific fixed-effects. In the partial pooling with random-effects, both parameters $E_{max}^{(i)}$ and $ED_{50}^{(i)}$ are assumed to be schedule-specific random-effects. The simulation results are listed in Table 3. In the scenario of $E_{max}^{monthly} = -60\%$, the complete pooling approaches result in lower MAE in comparison to the partial pooling approaches, while reaching the coverage probability of 95% for the confidence intervals. However, in other scenarios, complete pooling approaches yield worse performance in terms of the MAE and coverage probabilities compared to the partial pooling approaches. The partial pooling with random-effects results in smaller MAE and the shorter credible intervals compared to the partial pooling with fixed-effects in all three scenarios.

When we take into account all simulation results, the partial pooling approaches are more robust in terms of the MAE and the coverage probabilities across scenarios compared to the complete pooling approaches. The partial pooling with random-effects yields better performance than the partial pooling with fixed-effects in terms of the MAE and the mean length of the credible intervals with the exception of highly heterogeneous scenarios.

5 Revisiting the Dupilumab trial

We return to the dupilumab trial which was described in Section 2. The least square means and standard errors for different arms of the trial are listed in Table 4 as reported in Thaçi et al. (2016). In total, 379 patients completed the trial. We analyzed the dataset assuming normal distribution for least square means with the given standard errors. Note that this is different than assuming normality for the observations as described in Equation (1) as reported in the reference paper for convenience (Thaçi et al., 2016). This will show that the proposed method also works with weaker assumption, as we only use an arm-level data instead of an observation-level data. Five different models were fitted in a Bayesian framework. We compare them via the approximate leave-one-out cross-validation information criteria (LOO-IC) (Vehtari et al., 2017). Note that LOO-IC has the same purpose as the Akaike Information Criteria (AIC) used in the frequentist framework and similar to the AIC, the lower value indicates the better model. All models assume an Emax model for the dose-response relationship. We use prior distributions described in Section 3.2. The model descriptions are listed in Table 5. Model 1 corresponds to the complete pooling. In Models 2-5, the $E_0^{(i)}$ are assumed to be shared between schedules, while $ED_{50}^{(i)}$ and $E_{max}^{(i)}$ are treated differently in each model. Hence, Models 2-5 are partial pooling approaches. In Models 2 and 3, $E_{max}^{(i)}$ are assumed to be shared between schedules. Model 2 assumes schedule-specific fixed-effects for $ED_{50}^{(i)}$, while Model 3 uses schedule-specific random-effects for $ED_{50}^{(i)}$. Model 4 assumes schedule-specific fixed-effects both for $ED_{50}^{(i)}$ and $E_{max}^{(i)}$, whereas Model 5 uses schedule-specific random-effects both for $ED_{50}^{(i)}$ and $E_{max}^{(i)}$. For the complete pooling, the doses are transformed into the biweekly scale, thus the new set of doses are $\{0, 50, 150, 200, 300, 600\}$. For

Table 3: Simulation results for varying $E_{max}^{monthly}$ scenarios. The mean absolute error (MAE) for the dose-response function, coverage probabilities and mean length of the interval estimates for the dose-response function obtained by the four methods. Four methods include complete pooling approaches using frequentist, CP (Frequentist), and Bayesian methods, CP (Bayesian), and partial pooling approaches using schedule-specific fixed-effects (PP - FE) and schedule-specific random-effects (PP - RE) for $ED_{50}^{(i)}$.

	$E_{max}^{monthly}$		
	-60%	-70%	-50%
Mean absolute error			
CP (Frequentist)	1.63	2.44	2.51
CP (Bayesian)	1.62	2.23	2.87
PP - FE	2.04	2.03	2.03
PP - RE	1.88	1.90	2.02
Coverage probability			
CP (Frequentist)	0.95	0.87	0.84
CP (Bayesian)	0.95	0.87	0.84
PP - FE	0.96	0.96	0.96
PP - RE	0.96	0.96	0.95
Mean length			
CP (Frequentist)	5.78	5.77	5.82
CP (Bayesian)	5.53	5.48	5.47
PP - FE	6.96	6.96	6.95
PP - RE	6.59	6.64	6.71

Models 3 and 5, we use the biweekly schedule as the reference schedule.

Table 5 displays the LOO-IC values for the five models. The complete pooling results in the best model in terms of the LOO-IC. The second and third best models are the partial pooling with schedule-specific random-effects for $ED_{50}^{(i)}$ and the partial pooling with schedule-specific fixed-effects for $ED_{50}^{(i)}$, respectively. Apparently, the model complexity is heavily penalized by LOO-IC for this dataset, hence LOO-IC results in lower values for the simpler models. One possible reason is the data sparsity, numbers of dose levels available for different schedules are 2, 3, and 3 (by including placebo arm for all schedules). Based on these results, hereafter, we focus on Models 1-3.

The posterior estimates obtained by Model 1 (Complete Pooling), Model 2 (PP - FE), and Model 3 (PP - RE) are shown in Table 6. Recall that for Models 2 and 3, the E_{max} parameters are shared between schedules. The estimates ED_{50}^{weekly} and $ED_{50}^{monthly}$ of the complete pooling are calculated by re-scaling the estimate of $ED_{50}^{biweekly}$. Across three methods, estimates of E_0 are quite similar. For E_{max} and $ED_{50}^{(i)}$, however, the partial pooling with fixed-effects yields different results compared to the complete pooling and the partial pooling with random-effects. The heterogeneity parameter $\tau_{ED_{50}}$ results in high uncertainty (posterior mean 0.5 with standard deviation of 0.5), indicating the complete pooling is adequate. The estimated dose-response functions $\hat{f}$ by the complete pooling, the partial pooling with fixed-effects, and the partial pooling with random-effects are displayed in Figure 4. The $\hat{f}(t)$ are the posterior medians for the dose-response function $f(t)$ evaluated at each i where $i \in \{0, 20.7, \ldots, 600$ (mg/m^2-biweekly)$\}$, equidistant sequence between 0 and 600 with 30 elements.

Table 4: The dupilumab trial: Sample sizes, least square (LS) means, and standard errors for each arm in the trial.

Arm	Schedule	Dose (mg/m^2)	Sample size	LS mean	Standard error
1	Weekly	0	61	-18.1	5.2
2	Weekly	300	63	-73.7	5.2
3	Biweekly	200	61	-65.4	5.2
4	Biweekly	300	64	-68.2	5.1
5	Monthly	100	65	-44.8	5.0
6	Monthly	300	65	-63.5	4.9

Table 5: Analyzing the dupilumab trial: The approximate leave-one-out information criterion (LOO-IC) obtained by five different models. In all models, $E_0^{(i)}$ are assumed to be shared between schedules. The first model is the complete pooling, thus effectively all model parameters are assumed to be shared.

Model	$ED_{50}^{(i)}$	$E_{max}^{(i)}$	LOO-IC
Model 1	Shared	Shared	36.0
Model 2	Fixed-effects	Shared	39.8
Model 3	Random-effects	Shared	37.4
Model 4	Fixed-effects	Fixed-effects	41.7
Model 5	Random-effects	Random-effects	41.1

Similarly, 95% equi-tailed credible intervals evaluated at each i are displayed in Figure 4. The median dose-response curve obtained by the complete pooling and the partial pooling with random-effects are very similar, which is in alignment with the posterior estimates shown in Table 6. The median dose-response curve estimated by the partial pooling with fixed-effects is slightly different from the complete pooling and the partial pooling with random-effects. As expected, the complete pooling produces the shortest 95% credible intervals around $\hat{f}$, whereas the partial pooling with fixed-effects gives the widest. Such behaviour was also observed in the simulations. The dupilumab trial is similar to the scenarios when the sample size for each arm is 60, and both $ED_{50}^{\text{biweekly}}$ and ED_{50}^{monthly} do not deviate much from $ED_{50}^{\text{biweekly}}$, meaning that low heterogeneity in $ED_{50}^{(i)}$ between schedules. Additionally, Figure 5 (Appendix B) demonstrates the marginal posterior density estimates of $ED_{50}^{(i)}$ obtained by three methods alongside with the priors used for $ED_{50}^{(i)}$ in the partial pooling with fixed-effects. The posterior and prior distribution for the $ED_{50}^{\text{biweekly}}$ parameter are very similar in the partial pooling with fixed-effects. Recall that other than the placebo arm, there is only one arm with weekly schedule, hence indicating the data sparsity problem. In conclusion, although the complete pooling may be sufficient for this particular application, we obtain very similar dose-response estimates by using the partial pooling with random-effects.

6 Conclusions and outlook

An assumption of the homogeneity between schedules can be considered unrealistic, hence a partial pooling is more reasonable than the complete pooling. Rather than using schedule-specific fixed-effects in a partial pooling approach, we have proposed to use schedule-specific fixed-effects for the certain

Table 6: The estimates obtained by analyzing the dupilumab trial. Posterior means and standard deviations obtained by the complete pooling, the partial pooling with fixed-effects (PP - FE), and the partial pooling with random-effects (PP - RE) are shown. See the main text for the descriptions of the methods.

	CP		PP - FE		PP - RE	
	Mean	SD	Mean	SD	Mean	SD
E_0	-18.5	4.9	-18.1	5.0	-18.2	5.1
E_{max}	-61.0	7.4	-56.9	8.0	-60.0	8.6
ED_{50}^{weekly}	32.3	15.1	20.4	27.0	30.0	29.2
$ED_{50}^{biweekly}$	64.6	30.3	37.4	35.3	56.9	40.6
$ED_{50}^{monthly}$	129.1	60.6	100.0	46.2	116.7	58.7
$\tau_{ED_{50}}$					0.5	0.5

parameters such as ED_{50}, allowing dynamically borrowing information in a fully Bayesian framework. In simulation studies, the proposed method displayed more robust performance in terms of the mean absolute error and coverage probabilities for the dose-response function $f(t)$ compared to the complete pooling. Furthermore, the proposed method produces lower mean absolute error and shorter interval estimates for $f(t)$ across most of the scenarios compared to using schedule-specific fixed-effects in a partial pooling approach.

In this paper, we focused on the Emax model for the dose-response function. To account for the model uncertainty, it is important to consider alternative functions, such as log-linear or exponential. The shrinkage estimation can be applied to such alternative dose-response models, as well. One way of dealing with the model uncertainty is using a model selection criteria (e.g. AIC in the frequentist context) to decide the right functional form. Hence, by using a criteria such as LOO-IC, one can utilize the proposed approach to analyze data from a phase II trial with multiple schedules. Alternatively, a Bayesian model averaging approach (Schorning et al., 2016) can be used to deal with uncertainty of dose-response models. Here, we consider phase II trials with multiple schedules. Instead of multiple schedules, one may investigate phase II trials with multiple subgroups, for example multiple patient populations. The proposed method is still applicable for such situations.

The parametrization used in the proposed method, Equation (3), can be considered hard to motivate, since an overall mean of schedule-specific estimates does not have a meaningful interpretation. This can be overcome by adopting an asymmetric parametrization of schedule-specific estimates in terms of a reference schedule as follows

$$\begin{aligned} ED_{50}^{(k^*)} &\sim \mathcal{N}(\alpha_{ED50}, 0) \quad (\text{i.e. } ED_{50}^{(k^*)} = \alpha_{ED50}) \\ ED_{50}^{(k)} &\sim \mathcal{N}(\alpha_{ED50}, \beta_{ED50}^2) \end{aligned}$$

where α_{ED50} and β_{ED50} are the location and scale parameters, respectively (Röver and Friede, 2020).

Although the partial pooling with random-effects is an improvement to complete pooling and the partial pooling with fixed-effects, the exchangeability assumption bears the risk of too much shrinkage. Perhaps, it is not very desirable to allow borrowing information for the extreme schedule. To overcome this, the exchangeability-nonexchangeability (EXNEX) models (Neuenschwander et al., 2016) can be considered. EXNEX models can be used to share information across similar schedules, while avoid too much borrowing for the extreme schedule. However, such complicated models should be calibrated well, due to sparse data available in a typical phase II trial.

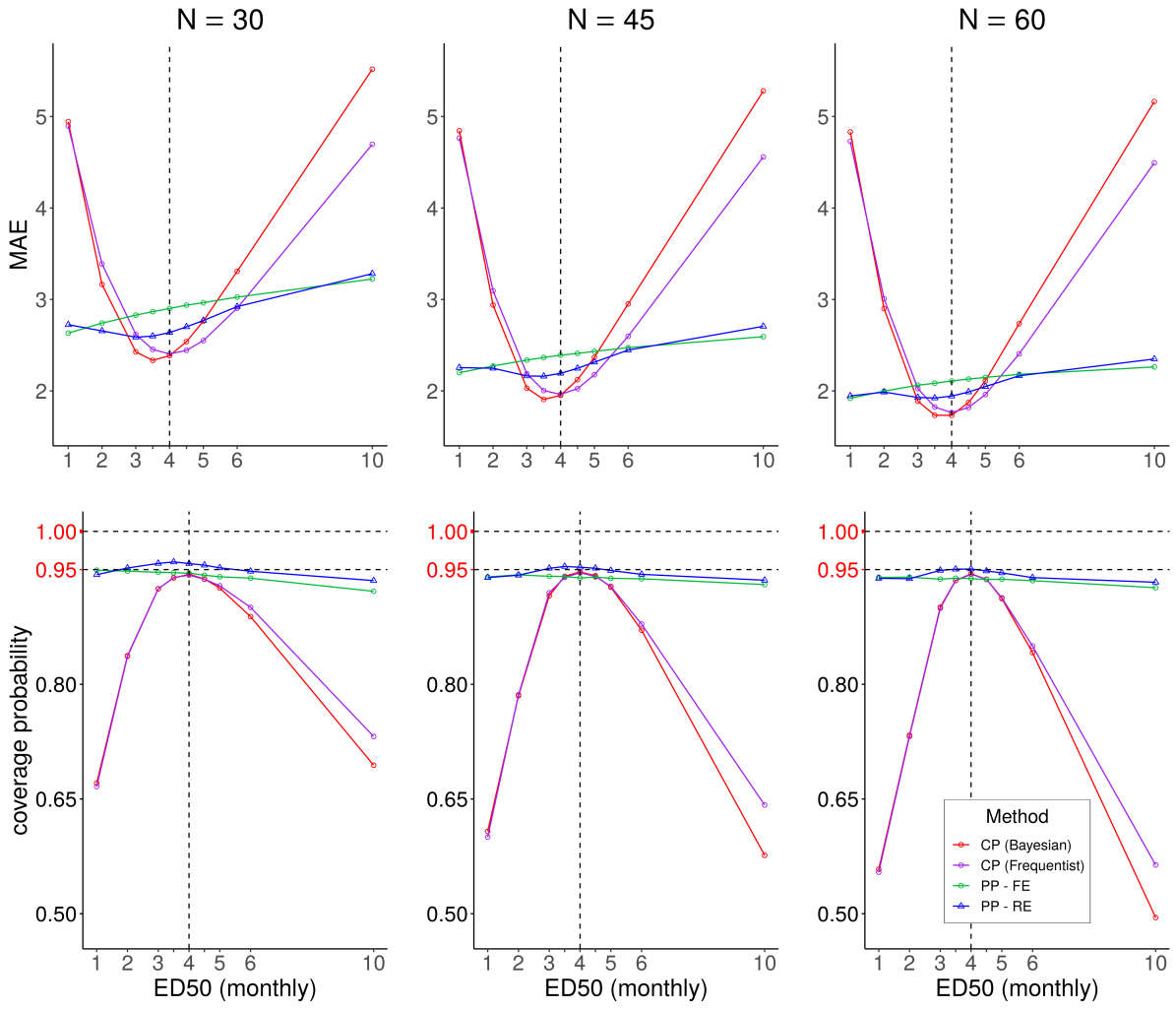

Figure 2: Simulation results for different sample sizes H per arm. The mean absolute error (MAE) and coverage probabilities for the dose-response curve obtained by four methods with different sample sizes. Four methods include complete pooling approaches using frequentist, CP (Frequentist), and Bayesian methods, CP (Bayesian), and partial pooling approaches using schedule-specific fixed-effects (PP - FE) and schedule-specific random-effects (PP - RE) for $\text{ED}_{50}^{(i)}$. The vertical dashed line indicates the scenario without heterogeneity in the re-scaled $\text{ED}_{50}^{(i)}$ between biweekly and monthly schedules. $\text{ED}_{50}^{\text{biweekly}}$ is assumed to be 2 mg/kg.

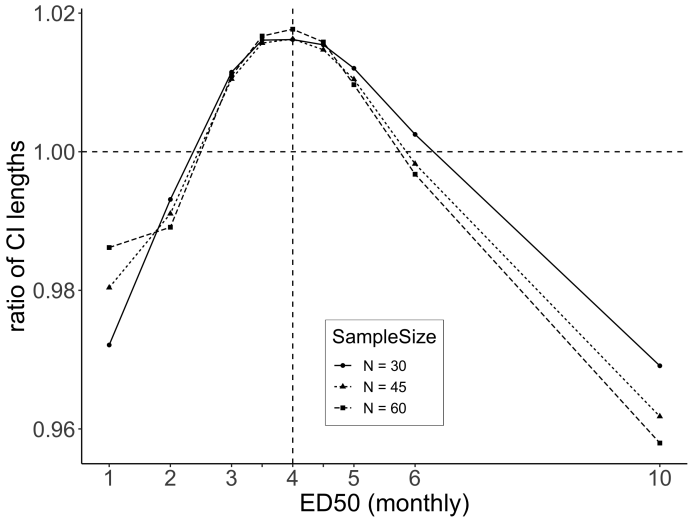

Figure 3: Simulation results for different sample sizes. Ratios of lengths of credible intervals for the dose-response curves obtained by the partial pooling with random-effects and the partial pooling with fixed-effects with different sample sizes. The denominator of the ratio is the length of credible interval obtained by the partial pooling with random-effects. The vertical dashed line indicates the scenario without heterogeneity in the re-scaled $\mathrm{ED}_{50}^{(i)}$ between biweekly and monthly schedules. $\mathrm{ED}_{50}^{\mathrm{biweekly}}$ is assumed to be 2 mg/kg.

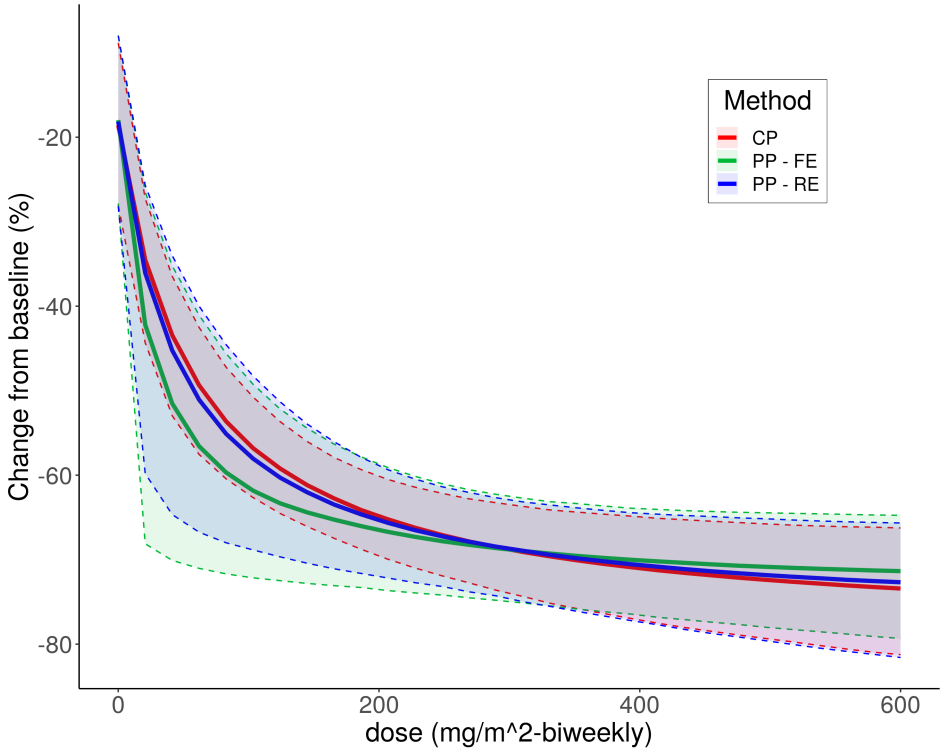

Figure 4: Dose-response curve and credible intervals for biweekly schedule obtained by the complete pooling (CP), the partial pooling with fixed-effects (PP - FE), and the partial pooling with random-effects (PP - RE) are shown. See the main text for the descriptions of the methods.

Acknowledgements

We are grateful to Monika Jelizarow and Christian Röver who contributed valuable comments and pointed us to several important references.

Conflict of interest

P.M. is an employee of Galapagos NV. T.F. is a consultant to Galapagos NV. MOR106 was jointly discovered by Galapagos NV and MorphoSys, and an MOR106 trial was used to motivate and illustrate the investigations presented here.

A How to use the ModStan R package?

The development version of `ModStan` is available on Github (https://github.com/gunhanb/ModStan) and can be installed as follows:

```
library("devtools")
install_github("gunhanb/ModStan")
```

The dupilumab trial described in the text is available in the package, and it can be loaded as follows:

```
library("ModStan")
data("dat.Dupilumab")
```

See `?dat.Dupilumab` for the description of the dataset.

The `mod_stan` is the main fitting function of the package. The main computations are executed via the `rstan` package's `sampling` function. We can fit the partial pooling method with schedule-specific random-effects for the $\text{ED}_{50}^{(i)}$ parameter as follows:

```
PP.RE.Dupilumab.stan = mod_stan(dose = dose,
                                resp = resp,
                                sigma = sigma,
                                schedule = schedule,
                                freq = freq,
                                freq_ref = 24 * 7 * 8,
                                data = dat.Dupilumab,
                                model = "PP-RE",
                                tau_prior_dist = "half-normal",
                                tau_prior = 1,
                                chains = 3,
                                stan_seed = 111,
                                iter = 4000,
                                warmup = 2000)
```

Convergence diagnostics and the results can be very conveniently obtained using the `shinystan` package as follows:

```
library("shinystan")
launch_shinystan(as.shinystan(PP.RE.Dupilumab.stan$fit))
```

The posterior summary statistics can be obtained using the following command:

```
PP.RE.Dupilumab.stan
```

B Marginal posterior density estimates of ED_{50} (dupilumab trial)

The marginal posterior density estimates obtained by the three methods (CP, PP - FE, PP - RE) are demonstrated in Figure 5. Also, the prior distribution used for the PP - FE is shown.

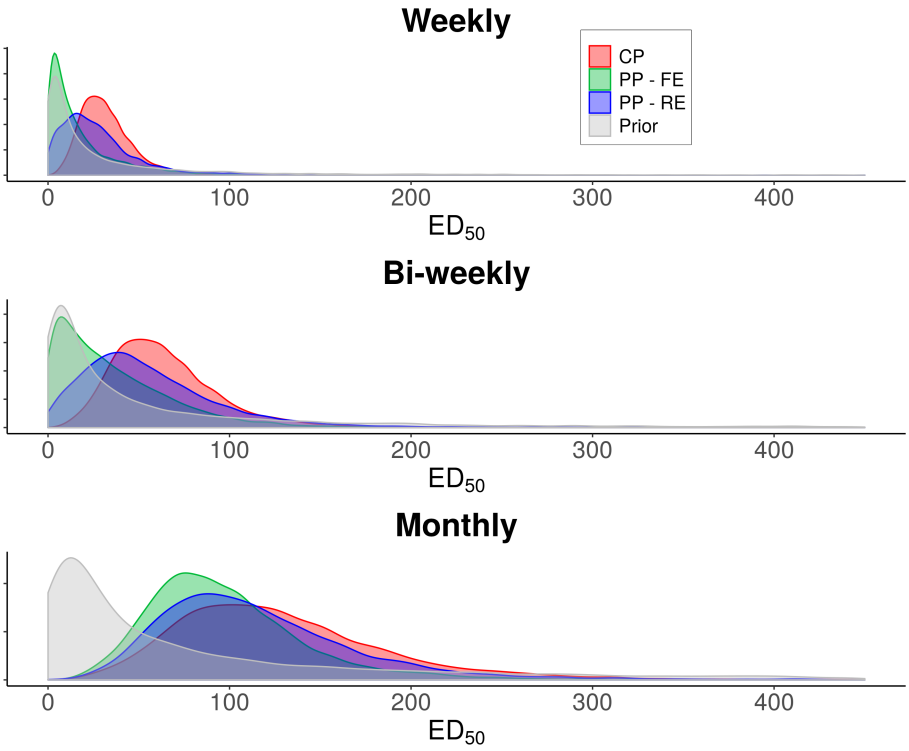

Figure 5: Marginal posterior density estimates of ED_{50} for weekly, biweekly, and monthly schedule obtained by the CP, the PP - FE, and the PP - RE. Also, shown is prior distributions used for the PP - FE.

References

Alexander, H., Patton, T., Jabbar-Lopez, Z., Manca, A., and Flohr, C. (2019). Novel systemic therapies in atopic dermatitis: what do we need to fulfill the promise of a treatment revolution? *F1000Research*, 8(132).

Betancourt, M. and Girolami, M. (2015). *Current trends in Bayesian methodology with applications*, chapter 4, pages 79–103. CRC Press, Boca Raton.

Bornkamp, B. (2012). Functional uniform priors for nonlinear modeling. *Biometrics*, 68(3):893–901.

Bornkamp, B. (2014). Practical considerations for using functional uniform prior distributions for dose-response estimation in clinical trials. *Biometrical Journal*, 56(6):947–962.

Bornkamp, B., Pinheiro, J., and Bretz, F. (2018). *DoseFinding: Planning and analyzing dose finding experiments*. R package version 0.9-16.

Bretz, F., Pinheiro, J., and Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61(3):738–748.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319.

Eichenfield, L. and Stein Gold, L. (2017). Systemic therapy of atopic dermatitis: Welcome to the revolution. *Seminars in cutaneous medicine and surgery*, 36(4S):S103–S105.

Feller, C., Schorning, K., Dette, H., Bermann, G., and Bornkamp, B. (2017). Optimal designs for dose response curves with common parameters. *Annals of Statistics*, 45(5):2102–2132.

Freidlin, B. and Korn, E. (2013). Borrowing information across subgroups in phase II trials: Is it useful? *Clinical Cancer Research*, 19(6):1326–1334.

Friede, T., Röver, C., Wandel, S., and Neuenschwander, B. (2017). Meta-analysis of few small studies in orphan diseases. *Research Synthesis Methods*, 8(1):79–91.

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

Giugliano, R., Desai, N., Kohli, P., Rogers, W., Somaratne, R., Huang, F., Liu, T., Mohanavelu, S., Hoffman, E., McDonald, S., Abrahamsen, T., Wasserman, S., Scott, R., and Sabatine, M. (2012). Efficacy, safety, and tolerability of a monoclonal antibody to proprotein convertase subtilisin/kexin type 9 in combination with a statin in patients with hypercholesterolaemia (LAPLACE-TIMI 57): a randomised, placebo-controlled, dose-ranging, phase 2 study. *The Lancet*, 380(9858):2007–2017.

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology*, 29(1):158–167.

Günhan, B., Röver, C., and Friede, T. (2020). Random-effects meta-analysis of few studies involving rare events. *Research Synthesis Methods*, 11(1):74–90.

Jones, H., Ohlssen, D., Neuenschwander, B., Racine, A., and Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. *Clinical Trials*, 8(2):129–143.

Mayo Clinic (2018). Atopic Dermatitis. https://www.mayoclinic.org/diseases-conditions/atopic-dermatitis-eczema/symptoms-causes/syc-20353273. Updated March, 2018. Accessed January, 2020.

Möllenhoff, K., Bretz, F., and Dette, H. (2019). Equivalence of regression curves sharing common parameters. *Biometrics*, pages 1–12.

MorphoSys AG (2019). MOR106 clinical development in atopic dermatitis stopped. https://www.morphosys.com/media-investors/media-center/morphosys-ag-mor106-clinical-development-in-atopic-dermatitis-stopped. Updated October, 2019. Accessed January, 2020.

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

Röver, C. and Friede, T. (2020). Dynamically borrowing strength from another study through shrinkage estimation. *Statistical Methods in Medical Research*, 29(1):293–308.

Ruberg, S. (1995). Dose response studies i. some design considerations. *Journal of Biopharmaceutical Statistics*, 5(1):1–14.

Schorning, K., Bornkamp, B., Bretz, F., and Dette, H. (2016). Model selection versus model averaging in dose finding studies. *Statistics in Medicine*, 35(22):4021–4040.

Spiegelhalter, D., Abrams, K., and Myles, J. P. d. (2004). *Bayesian Approaches to clinical trials and health-care evaluation.* West Sussex: CRC Press.

Thaçi, D., Simpson, E., Beck, L., Bieber, T., Blauvelt, A., Papp, K., Soong, W., Worm, M., Szepietowski, J., Sofen, H., et al. (2016). Efficacy and safety of dupilumab in adults with moderate-to-severe atopic dermatitis inadequately controlled by topical treatments: a randomised, placebo-controlled, dose-ranging phase 2b trial. *Lancet*, 387(10013):40–52.

Thomas, N., Sweeney, K., and Somayaji, V. (2014). Meta-analysis of clinical dose-response in a large drug development portfolio. *Statistics in Biopharmaceutical Research*, 6(4):302–317.

Varadhan, R. (2015). *alabama: Constrained nonlinear optimization.* R package version 2015.3-1.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., 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.

A.4 Random-effects meta-analysis of few studies involving rare events

Published version is publicly available from https://doi.org/10.1002/jrsm.1370.

Received: 12 September 2018 | Revised: 4 July 2019 | Accepted: 5 July 2019

DOI: 10.1002/jrsm.1370

RESEARCH ARTICLE

WILEY Research Synthesis Methods

Random-effects meta-analysis of few studies involving rare events

Burak Kürsad Günhan | **Christian Röver** | **Tim Friede**

Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

Correspondence
Burak Kürsad Günhan, Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany.
Email: burak.gunhan@med.uni-goettingen.de

Meta-analyses of clinical trials targeting rare events face particular challenges when the data lack adequate numbers of events for all treatment arms. Especially when the number of studies is low, standard random-effects meta-analysis methods can lead to serious distortions because of such data sparsity. To overcome this, we suggest the use of *weakly informative priors* (WIPs) for the treatment effect parameter of a Bayesian meta-analysis model, which may also be seen as a form of penalization. As a data model, we use a binomial-normal hierarchical model (BNHM) that does not require continuity corrections in case of zero counts in one or both arms. We suggest a normal prior for the log-odds ratio with mean 0 and standard deviation 2.82, which is motivated (a) as a symmetric prior centered around unity and constraining the odds ratio within a range from 1/250 to 250 with 95% probability and (b) as consistent with empirically observed effect estimates from a set of 37 773 meta-analyses from the Cochrane Database of Systematic Reviews. In a simulation study with rare events and few studies, our BNHM with a WIP outperformed a Bayesian method without a WIP and a maximum likelihood estimator in terms of smaller bias and shorter interval estimates with similar coverage. Furthermore, the methods are illustrated by a systematic review in immunosuppression of rare safety events following pediatric transplantation. A publicly available **R** package, `MetaStan`, is developed to automate a Bayesian implementation of meta-analysis models using WIPs.

KEYWORDS
Bayes, few studies, random-effects meta-analysis, rare events, weakly informative priors

1 | INTRODUCTION

Individual clinical studies are often underpowered to detect difference of probabilities or rates of rare events, for example, safety events, and thus, meta-analysis may be the only way to obtain reliable evidence of treatment differences with regard to the rare events.[1] On the other hand, meta-analysis of clinical studies for rare events faces particular challenges, since the numbers of events might be very small in some treatment arms. The problem is even more pronounced when some studies have no events either in one or in both treatment arms (so-called single-zero or double-zero studies).

The exclusion of the double-zero studies from the analysis can bias the treatment effect parameter estimate away from the null (especially for the unbalanced design)[2] and also causes loss of information, since double-zero studies contain information through their sample sizes.[3] Hence, we consider methods that do not remove double-zero studies from the analysis. Two established fixed-effect

meta-analysis methods exist for rare events, namely, Peto's method[4] and the Mantel-Haenszel (MH) method.[5] On the other hand, an assumption of homogeneity, that is, a single common parameter for all studies, is typically unrealistic for studies in the biomedical sciences.[6-8] Therefore, we focus on random-effects methods in this paper.

Standard (approximate) random-effects meta-analysis methods, for example, the normal-normal hierarchical model,[9] require a *continuity correction* in case of single-zero or double-zero studies, that is, the addition of a fixed value (typically 0.5) to all cells of the contingency table for studies with no events or with 100% events (no nonevents). Such simple approaches have been found problematic for meta-analyses involving rare events.[10] Therefore, statistical models based on exact distributional assumptions have been suggested. These include different parametrizations of the binomial-normal hierarchical model (BNHM),[11] a mixed effects conditional logistic model,[12] a Poisson-normal hierarchical model,[13] a Poisson-Gamma hierarchical model,[14] or a beta-binomial model (BBM).[3] In this paper, we focus on a parametrization of the BNHM that was suggested by Smith et al.[15]

Consider an extreme case of meta-analysis of rare events, where all studies include no events for the same treatment arm. These data sparsity problem in a meta-analysis can be seen as a *separation* problem in the logistic regression context[16] in which case a maximum likelihood estimate (MLE) for the treatment effect parameter does not exist. A very useful way to deal with separation problems, or, more generally, data sparsity in logistic regression is *penalization*, that is, adding a penalty (adjustment) term to the original likelihood function to regularize (or stabilize) the estimates.[17] In a frequentist framework, penalty terms may be specified so that these *nudge* the MLE into a desired direction if the maximum is not or poorly defined; one such example is Firth penalization.[17-19] From a Bayesian viewpoint, penalization may often be motivated as *weakly informative priors* (WIPs) that are multiplied to the likelihood function.[20]

Numbers of studies included in meta-analyses are typically small, posing additional challenges.[21] For Bayesian meta-analysis of few studies, different WIPs have been suggested for the heterogeneity parameter; see Chung et al[22] for penalized MLE approach and also see Gelman[23] and Friede and Röver[24] for Bayesian inference. Here, we consider the meta-analysis of few studies targeting rare events. To deal with data sparsity present in the meta-analysis of few studies with rare events, we suggest the use of WIPs for the treatment effect parameter in a fully Bayesian context inspired by penalization ideas.[17,20] We use a BNHM that is parameterized in terms of baseline risks and a treatment effect for the data. Note that this is a contrast-based model meaning that relative treatment effects are assumed to be exchangeable across trials.[25] Our suggested default WIP for the treatment effect parameter is motivated via the consideration of the prior expected range of treatment effect values. Furthermore, it is consistent with effect estimates empirically observed in a large set of meta-analyses from the Cochrane Database of Systematic Reviews (CDSR) with binary endpoints.

The main contribution of this paper is the introduction of default WIPs as penalization for treatment effect parameters to deal with data sparsity in the meta-analysis of few studies involving rare events. Another contribution is the introduction of an **R** package, **MetaStan** (https://CRAN.R-project.org/package=MetaStan), which is developed to automate a Bayesian implementation of meta-analysis models using WIPs as described in the paper and which is publicly available from CRAN. In Section 2, we describe a systematic review concerning rare safety events associated with immunosuppressive therapy following pediatric transplantation. In Section 3, we describe the application of WIPs for the treatment effect parameter. We review a BNHM for meta-analysis, discuss the derivation of WIP, and an empirical investigation of treatment effect parameter estimates from the CDSR. Long-run properties of different methods including the proposed one are investigated in the simulation studies in Section 5. In Section 6, the example is revisited to illustrate the proposed method and its implementation. We close with some conclusions and provide a discussion.

2 | AN APPLICATION IN PEDIATRIC TRANSPLANTATION

Several rare pediatric liver diseases can nowadays be successfully treated by liver transplantation with good long-term outcomes. Crins et al[26] conducted a systematic review of controlled but not necessarily random-

TABLE 1 Data on patient deaths and posttransplant lymphoproliferative disease (PTLD) from the meta-analysis in pediatric transplantation conducted by Crins et al[26]

	Outcome: Death				Outcome: PTLD			
	Control		Experimental		Control		Experimental	
	Events	Total	Events	Total	Events	Total	Events	Total
Heffron et al[29]	3	20	4	61	-	-	-	-
Schuller et al[30]	-	-	-	-	0	12	0	18
Ganschow et al[31]	3	54	1	54	0	54	1	54
Spada et al[32]	3	36	4	36	1	36	1	36
Gras et al[33]	3	34	2	50	-	-	-	-

ized studies of the Interleukin-2 receptor antibodies (IL-2RA) basiliximab and daclizumab in pediatric liver transplantation. Primary outcomes were acute rejections (ARs), steroid-resistant rejections (SRRs), graft loss, and death. Their analyses were based on a random-effects meta-analysis using a restricted maximum likelihood approach(REML).[27] Crins et al[26] used risk ratios as effect measures, while we use the odds ratios here. With rare events, however, these should be very similar. Heterogeneity was assessed using Cochrane's Q test.[28] Secondary outcomes included renal dysfunction nt lymphoproliferative disease (PTLD). For illustrative purposes, here, we focus on death and PTLDs, and these outcomes are displayed in Table 1.

The specific problems with meta-analyses concerning rare events outlined in the introduction are prominent here. Firstly, the numbers of events are very small. For the PTLD dataset, there is one single-zero study and one double-zero study. Secondly, there are few studies available, only four for deaths and three for PTLD. Empirical event rates are lower in three of the four experimental groups in the data on patient deaths. For PTLD, the data appear inconclusive for Schuller et al[30] and Spada et al,[32] and only a single event observed in the experimental group suggests an increased risk in the study by Ganschow et al.[31]

3 | WIPS FOR THE TREATMENT EFFECT

In this section, we present the usage of WIPs for the treatment effect parameter to conduct random-effects meta-analysis of rare events with few studies. As a data model, we review a BNHM and then show how to derive a WIP for a treatment effect parameter. Then, empirical evidence obtained from the CDSR supporting the choice of WIPs is illustrated.

3.1 | Data model

The BNHM has been introduced by Smith et al.[15] In the BNHM, for each trial $i \in \{1, \ldots, k\}$ and treatment arm $j \in \{0, 1\}$, the event counts r_{ij} are modeled using a binomial distribution, that is, $r_{ij} \sim \text{Bin}(\pi_{ij}, n_{ij})$. The logit link is used to transform π_{ij} onto the log odds scale where effects can be assumed to be additive

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \mu_i + \theta_i x_{ij} \\ \theta_i &\sim \mathcal{N}(\theta, \tau^2), \end{aligned} \tag{1}$$

where x_{ij} is a treatment indicator, namely, $+0.5$ = experimental ($j = 1$) and -0.5 = control ($j = 0$). The μ_i are fixed effects denoting the baseline risks of the event in each study i, θ is the mean treatment effect, and τ is the heterogeneity in treatment effects between trials. The BNHM belongs to the family of generalized linear mixed models (GLMMs); this family also includes models for other types of data including continuous or count outcomes. It is important to note here that by treating the baseline risks μ_i as fixed effects, the analysis effectively stratifies the risk by study, as pooling of risks might compromise the studies' randomization. In this sense, it constitutes a contrast-based model.[25] Unlike the normal-normal hierarchical model, the BNHM does not rely on a normal approximation, since it builds on the binomial nature of the data directly.

The BNHM can be fitted using frequentist approaches, for example, via maximum likelihood estimation (MLE).[11] Alternatively, Bayesian methods are commonly used. In a fully Bayesian approach, prior distributions for parameters θ, μ_i, and τ need to be specified. Note that the parameter θ is on the log-odds ratio scale whereas μ_i are on the log-odds scale. Baseline risks (μ_i) may be seen as intercept parameters in a standard logistic regression model. For μ_i, we use a vague normal prior with mean 0 and standard deviation 10, following the recommendation by Gelman et al.[20] The prior choice for θ is our main focus and will be discussed in Section 3.2. The prior choice for the heterogeneity parameter τ, which is a standard deviation parameter, has gained much attention in the literature as discussed in the introduction. Friede et al[24] have shown that for meta-analysis of few studies, the use of WIPs for τ displays desirable long-run properties in comparison with frequentist alternatives. Following their suggestions, we use a half-normal prior with scale of 0.5 ($\mathcal{HN}(0.5)$) for τ which has the median of 0.337 with an upper 95% quantile of 0.98. Values for τ of 0.25, 0.5, 1, and 2 represent moderate, substantial, large, and very large heterogeneity.[34] Thus, a $\mathcal{HN}(0.5)$ prior captures heterogeneity values for log-OR typically seen in meta-analyses of log-ORs and will therefore be a sensible choice in many applications.

3.2 | Derivation of a WIP for the treatment effect

A common prior choice for the treatment effect parameter θ is a noninformative (vague) prior such as normal distribution with a large variance, for example, $\mathcal{N}(0, 100^2)$. One way of constructing a WIP works via consideration of the prior expected range of treatment effect values.[35] Before the derivation of the WIP for treatment effect parameter θ, recall that θ is on the log-odds ratio scale. Thus, a value of $\theta = 0$ means an odds ratio of 1, ie, *no effect*, and a value of $\theta = 1$ means that odds differ by a factor (ratio) of $\exp(1) = 2.7$.

We assume a symmetric prior centered around zero, implying equal probabilities for positive or negative

treatment effects. Symmetry then implies (on the log-odds ratio scale) that

$$P(\theta > q) = P(\theta < -q), \tag{2}$$

where (on the odds ratio scale)

$$\exp(-q) = \frac{1}{\exp(q)}. \tag{3}$$

The prior's scale parameter σ_{prior} then may be set such that a priori the odds ratio is with 95% probability confined to a certain range:

$$P(1/\delta < \exp(\theta) < \delta) = 95\%. \tag{4}$$

In case of a normal prior with standard deviation σ_{prior}, we can then simply specify

$$\sigma_{\text{prior}} = \frac{\log(\delta)}{1.96}. \tag{5}$$

We conservatively specify δ as 250, meaning that we consider it unlikely that the odds ratio will be larger than 250 or smaller than 1/250. By plugging in this number into (5), we obtain $\sigma_{\text{prior}} = 2.82$.

Another way to motivate the prior standard deviation is by using the idea of *unit information priors*.[36,37] When the treatment effect parameter is on the log-odds ratio scale (as in the BNHM), then the standard error is given by $\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$. Assuming equal allocation, a neutral effect, and equal counts of events and nonevents, we can simply set the table allocation to $a = b = c = d = \frac{N}{4}$. Therefore, if we (heuristically) reverse the argument, a prior for the log-odds ratio with zero mean and 2.82 standard deviation gives[37]

$$2.82 \approx \sqrt{8} = \sqrt{\frac{1}{\frac{N}{4}}+\frac{1}{\frac{N}{4}}+\frac{1}{\frac{N}{4}}+\frac{1}{\frac{N}{4}}}. \tag{6}$$

Hence, $N = 2$. In other words, in terms of prior's effective sample size, the choice of $\sigma_{\text{prior}} = 2.82$ is equivalent to adding two patients to the dataset. From this, it follows that $\mathcal{N}(0, 2.82^2)$ is a reasonable choice as a WIP for θ.

Note also the analogy between this WIP and commonly used continuity corrections: Zero entries in a contingency tables are commonly *fixed* by adding a correction term of 0.5 to each table cell of the single-zero or double-zero study, which also amounts to a total of two patients *added* to the data. This way of conducting continuity correction adds two patients to each single-zero or double-zero study, while the use of WIP is equivalent to adding two patients to the whole dataset.

3.3 | Empirical evidence supporting the WIP for the treatment effect

For an empirical investigation of the WIP for treatment effect parameter, we consider the meta-analysis datasets archived in the CDSR. All systematic reviews in the CDSR are available on the Cochrane Library website,[38] and personal or institutional access is required. For downloading the data from the CDSR and converting to CSV files, we use the program `Cochrane_scraper` (version 1.1.0).[39] We were able to access all Cochrane systematic reviews available in March 2018 (CD000004 to CD012788).

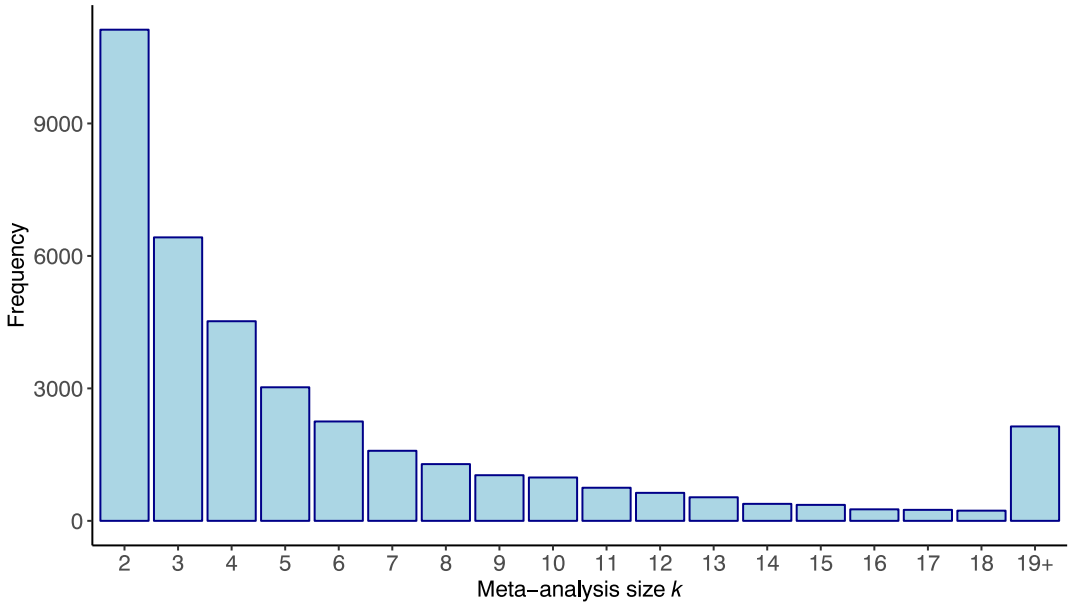

FIGURE 1 The distribution of numbers of studies included in each meta-analysis obtained from the Cochrane Database of Systematic Reviews (CDSR). The category labelled *19+* corresponds to meta analyses of size 19 or larger [Colour figure can be viewed at wileyonlinelibrary.com]

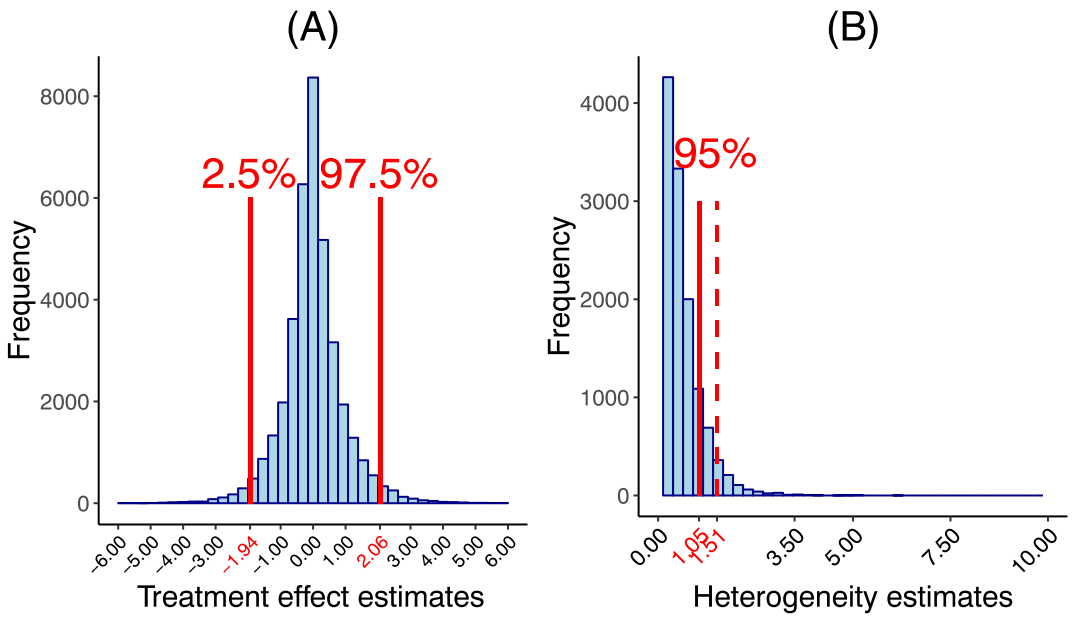

FIGURE 2 The distribution of the estimates of the mean treatment effect parameter θ A, and the distribution of the estimates of the (nonzero) heterogeneity standard deviation parameter τ B, obtained from the reanalysis of meta-analysis datasets in Cochrane Database of Systematic Reviews (CDSR) when the binomial-normal hierarchical model (BNHM) via maximum likelihood estimate (MLE) is used for estimation. In A, two red lines (−1.94 and 2.06) show the 2.5% and 97.5% quantiles of the θ estimates, respectively. In B, the solid red line (1.05) and the dashed red line (1.51) indicate the 95% quantiles of the τ estimates including zero-estimates and excluding zero-estimates, respectively. The fraction of zero-estimates of τ is 63% [Colour figure can be viewed at wileyonlinelibrary.com]

Meta-analyses were excluded if they included only one study, if the analysis was labelled as a subgroup or sensitivity analysis or there was insufficient information for classification, or if all data within the meta-analysis appeared to be erroneous. Finally, we only consider meta-analyses with dichotomous outcomes. In total, 37 773 meta-analysis datasets from 4712 reviews are included. Note that we did not distinguish regarding efficacy or safety analyses.

The frequency of the number of studies k considered for each meta-analysis is illustrated in Figure 1. The percentage of the meta-analyses including five or less studies is 66%. This figure is consistent with other re-analyses of the CDSR (see, eg, previous works[8,21,40]). We re-analyzed the meta-analysis datasets from the CDSR using the BNHM via an MLE approach. This procedure is implemented using the **R** package `lme4`.[41] A histogram of the estimates of θ is illustrated in Figure 2A; 2.5% and 97.5% quantiles of the estimates of θ are −1.94 and 2.06, respectively. By following Turner et al,[40] we exclude the zero heterogeneity estimates; nonzero estimates of τ are shown in Figure 2B. The fraction of nonzero heterogeneity estimates is 63%, which is also consistent with previous findings.[40] The 95% quantile of nonzero estimates of τ is 1.51, while the 95% quantile of τ estimates including zeroes is 1.05. The underlying distribution of the estimates of θ and τ and their variability are useful to see how large these estimates are in some general population, in this case the CDSR. Thus, these give us a rough sense of what would be a reasonable default prior distribution. Therefore, we suggest the use of WIPs, $\mathcal{N}(0, 2.82^2)$ for θ and $\mathcal{HN}(0.5)$ for τ, which are consistent with estimates of θ and τ empirically observed in the CDSR, meaning that both indicate odds ratios within reasonable ranges, and heterogeneity mostly below 1.0.

4 | IMPLEMENTATION OF THE PROPOSED PROCEDURE IN R USING STAN

The Bayesian implementation of the BNHM can be fitted with the probabilistic programming language **Stan**,[42] which employs a modern Markov chain Monte Carlo (MCMC) algorithm, namely, Hamiltonian Monte Carlo with the No-U-Turn Sampler.[43] It is known that the parametrization of the model can affect the performance of an MCMC algorithm. In the presence of sparse data such as in the meta-analysis of few studies involving rare events, Betancourt et al[44] showed that centered parametrization of a hierarchical model (such as the BNHM) brings computational issues compared with a noncentered parametrization. Thus, we use the noncentered reparametrized version of the BNHM for our implementations. Specifically, applying both location and scale reparametrization, (3.1) becomes $\mu_i + \theta x_{ij} + u_i \tau x_{ij}$ where $u_i \sim \mathcal{N}(0, 1)$ and $x_{ij} = +0.5$ (experimental) or $x_{ij} = -0.5$ (control). (Correction added on 06 January 2020, after first online publication: The preceding equation has been updated from "$\mu_i + \theta_i x_{ij} + u_i \tau^2$" to "$\mu_i + \theta x_{ij} + u_i \tau x_{ij}$".)

```
data {
  int<lower=1> N;                          // num studies
  int<lower=0> rctrl[N];                   // num events, control
  int<lower=1> nctrl[N];                   // num patients, control
  int<lower=0> rtrt[N];                    // num events, treatment
  int<lower=1> ntrt[N];                    // num patients, treatment
  vector[2] mu_prior;                      // Prior parameters for mu (mean and stdev)
  vector[2] theta_prior;                   // Prior parameters for theta (mean and stdev)
  real tau_prior;                          // Prior scale for tau
  int tau_prior_dist;                      // Indicator for prior distribution of tau
}

parameters {
  vector[N] mu;                            // baseline risks (log odds)
  real theta;                              // relative trt effect (log OR)
  vector[N] zeta;                          // individual trt effects
  real<lower=0> tau;                       // heterogeneity stdev.
}

transformed parameters {
  real pctrl[N];
  real ptrt[N];

  for(i in 1:N) {
    pctrl[i] = inv_logit(mu[i] - theta * 0.5 - zeta[i] * tau * 0.5);
    ptrt[i] = inv_logit(mu[i] + theta * 0.5 + zeta[i] * tau * 0.5);
  }
}

model {
  // latent variable (random-effects)
  zeta ~ normal(0, 1);
  // prior distributions
  mu ~ normal(mu_prior[1], mu_prior[2]);
  theta ~ normal(theta_prior[1], theta_prior[2]);
  if(tau_prior_dist == 1)  tau ~ normal(0, tau_prior)T[0,];
  if(tau_prior_dist == 2)  tau ~ uniform(0, tau_prior);
  if(tau_prior_dist == 3)  tau ~ cauchy(0, tau_prior)T[0,];
  // likelihood
  rctrl~ binomial(nctrl, pctrl);               // control event count
  rtrt ~ binomial(ntrt, ptrt);                 // treatment event count
}
```

Listing 1: **Stan** code defining the binomial-normal hierarchical model. (Correction added on 06 January 2020, after first online publication: Listing 1 has been updated)

For practical applications, learning **Stan**'s syntax and the required knowledge of available features in **Stan** might present a hurdle preventing application of **Stan**. To this end, we developed a new **R** package `MetaStan` which is a purpose-built package defined on top of `Rstan`, the **R** interface for **Stan**. Our package `MetaStan` (https://CRAN.R-project.org/package=MetaStan) includes the precompiled **Stan** model of the BNHM, which eliminates the compilation time and the need of learning **Stan**'s syntax. The **Stan** code for the BNHM is shown in Listing 1. `MetaStan` includes different options for WIPs of the model parameters of the BNHM. `MetaStan` syntax is similar to the syntax of the popular meta-analysis package `metafor`[27] so that it should be easy for a `metafor` user to utilize our package. The syntax of `MetaStan` is displayed for the pediatric transplantation example in Section 6, and in Appendix A, we show how to install and use `MetaStan`.

5 | SIMULATION STUDY

In order to assess the long-run properties of the proposed approach and compare it with some alternatives, we conducted a simulation study.

5.1 | Simulation setup

The simulation scenarios are similar to those considered by Friede et al,[24] but with the important difference that we focus on rare events. The datasets are generated under the BNHM, more specifically (3.1). Numbers of studies ($k \in \{2, 3, 5\}$) and true treatment effects ($\theta = \{-5, -4, -3, -2, -1, -0.5, 0, 0.5, 1, 2, 3, 4, 5\}$) are varied, resulting in a total of 39 simulation scenarios. To reflect the rare-event cases, true baseline risks on the probability scale are taken uniformly between 0.005 and 0.05. Following Kuss,[3] a log-normal distribution is fitted to the sample sizes obtained from the CDSR data, resulting in a log-normal distribution with parameters $\mu = 5$ and $\sigma = 1$. Hence, sample sizes are generated from $\mathcal{LN}(5, 1)$, the minimum sample size is restricted to two patients (values below 2 are rounded up to 2), and at least one patient in each treatment arm is assumed. The degree of heterogeneity (τ) is taken as $\tau = 0.28$ (moderate heterogeneity), which is the median value of the predictive distribution for between-study heterogeneity in a meta-analysis in a general setting as estimated by Turner et al.[40] According to a binomial probability of 0.5, patients were allocated to the treatment groups, thus mimicking randomization. The simulations were carried out with 10 000 replications per scenario. The data sparsity is reflected in the average fractions of single-zero or double-zero studies in a simulated meta-analysis dataset, which are shown in Figure 3A. Notice that the fractions of the single-zero and double-zero studies are the highest when true treatment effect is −5, and they are decreasing with the increase of the treatment effect.

The proposed approach (BNHM using a WIP, that is $\mathcal{N}(0, 2.82^2)$, for θ: *WIP*) and four comparators are included in the analysis, namely, BNHM using a vague prior ($\mathcal{N}(0, 100^2)$) for θ (*Vague*), BNHM using MLE (*MLE*), the Mantel-Haenszel (*MH*) method[5] and a Bayesian implementation of the beta-binomial model (*BBM*).[3] It is important to note the differences of the MH and BBM from the BNHM methods. MH is a fixed-effect meta-analysis method, and BBM has a different underlying data generating process than the BNHM. For both Vague and WIP approaches, the prior for τ and μ are taken as $\mathcal{HN}(0.5)$, and $\mathcal{N}(0, 10^2)$, respectively. The MH estimator of the treatment effect parameter is given by

$$\hat{\theta}_{\text{MH}} = \frac{\sum_{i=1}^{k} \frac{r_{i1}(n_{i0} - r_{i0})}{n_i}}{\sum_{i=1}^{k} \frac{r_{i0}(n_{i1} - r_{i1})}{n_i}},$$

where $n_i = n_{i0} + n_{i1}$. In the BBM, the event counts r_{ij} are modeled using a binomial distribution, $r_{ij} \sim \text{Bin}(\pi_{ij}, n_{ij})$, as in the BNHM. The probabilities of event are assumed to be beta distributed: $\pi_{ij} \sim \text{Beta}(\alpha_j, \beta_j)$ where both arms share the same correlation parameter $\rho = \frac{1}{\alpha_0+\beta_0+1} = \frac{1}{\alpha_1+\beta_1+1}$, implying $\alpha_0 + \beta_0 = \alpha_1 + \beta_1$. It is common to reparametrize the model using mean parameters Φ_j such that $\Phi_j = \frac{\alpha_j}{\alpha_j+\beta_j}$. Finally, the linear predictor can be written as $\text{logit}(\Phi_j) = \mu + \theta x_j$ where θ is the parameter for the treatment effect, and x_j is a treatment indicator, 1 = experimental ($j = 1$) and 0 = control ($j = 0$). Vague priors are chosen for all parameters, namely, uniform priors across the interval [0,1] for all three parameters: Φ_0, Φ_1, and ρ.

Three MCMC chains were run in parallel for a total of 2000 iterations including 1000 iterations of burn-in. These values are tested in some replications; convergence

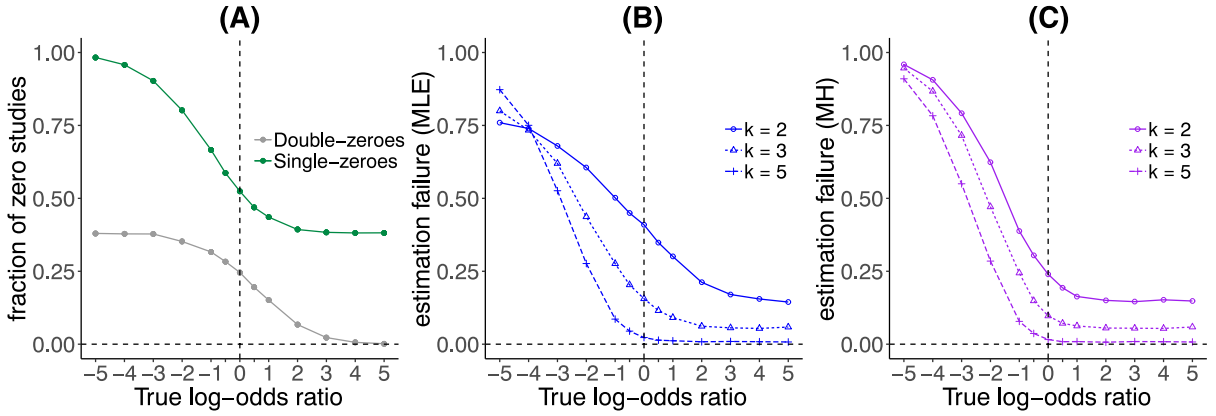

FIGURE 3 The average fraction of single-zero or double-zero studies in a simulated meta-analysis dataset A, and the fraction of the estimation failure for maximum likelihood estimate (MLE) and Mantel-Haenszel (MH) with different numbers of studies k used in the simulations (B and C) are shown [Colour figure can be viewed at wileyonlinelibrary.com]

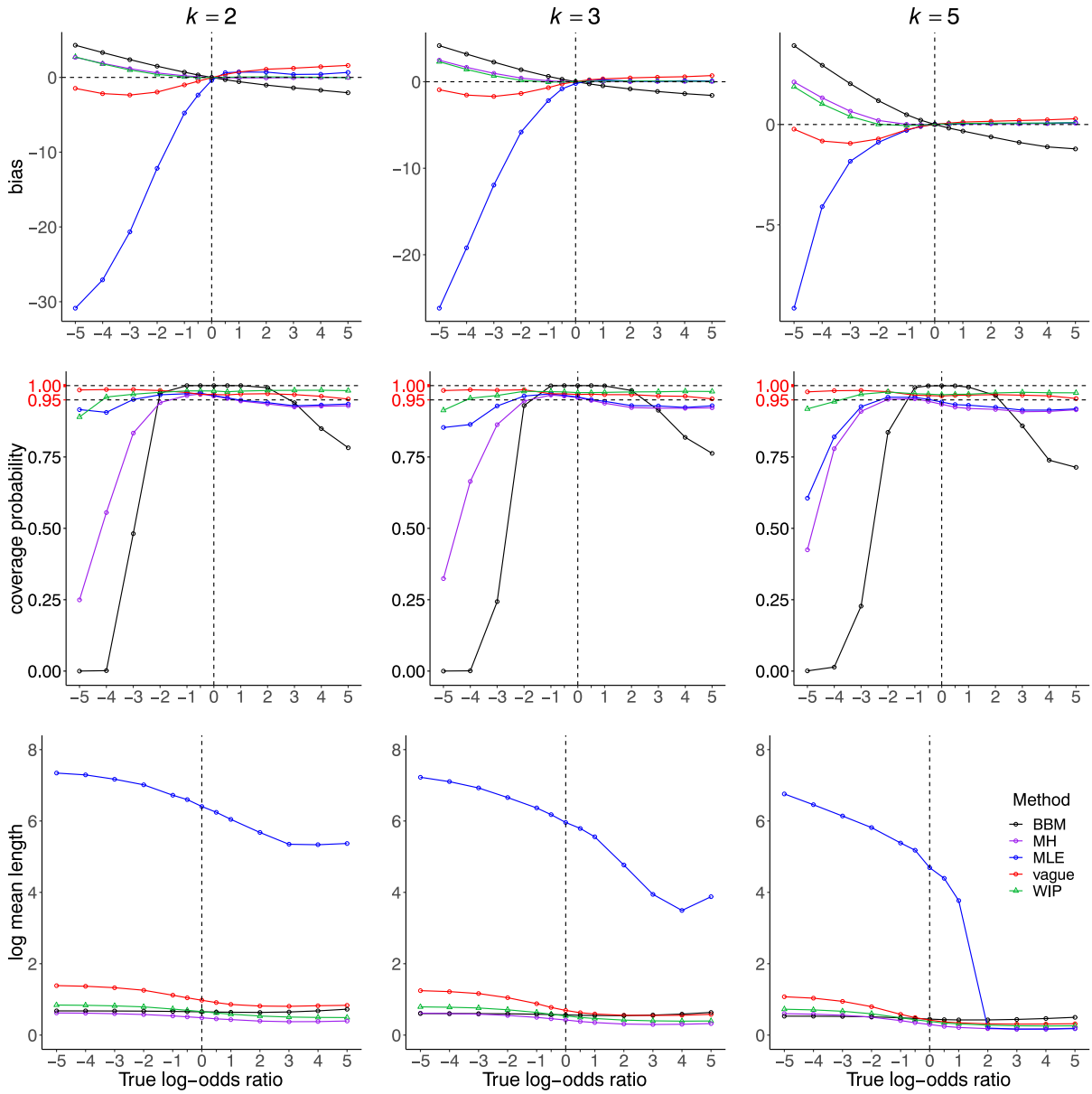

FIGURE 4 The bias for the mean treatment effect θ, coverage probabilities, and log mean length of the interval estimates for θ obtained by five methods (beta-binomial model [BBM], Mantel-Haenszel [MH], maximum likelihood estimate [MLE], Vague, and weakly informative prior [WIP]) are shown [Colour figure can be viewed at wileyonlinelibrary.com]

diagnostics are assessed and chosen accordingly. All chains were assumed to have reached convergence (no estimation failure). We used the package `lme4` for the MLE (using the adaptive Gauss-Hermite approximation to the maximum log-likelihood) and `metafor` for the MH (without using any continuity corrections) whereas the Vague, WIP, and BBM methods were fitted with our **MetaStan** package. Note that we use highest density intervals (HDI), which are the shortest credible intervals, as opposed to the commonly used equal-tailed credible intervals. The HDI were obtained using the HDInterval[45] package. All computations were performed using **R**.[46] The code for the computations for all methods used in the simulations is provided in Appendices A to D.

5.2 | Simulation results

For the MLE and the MH, the fractions of estimation failures are shown in Figure 3B and 3C. Estimation failure occurred for the MLE when the Gauss-Hermite approximation does not converge to the maximum log-likelihood, or when the MH estimator is not defined. The MLE and MH methods show very similar behavior of the estimation failure. Estimation failure is closely related to the fraction of meta-analysis datasets including single-zero or double-zero studies in the dataset, which can be seen by comparing Figure 3A and 3B,C. This is because when the data are highly sparse, estimation becomes more challenging for both MLE and MH. As a performance measure, we

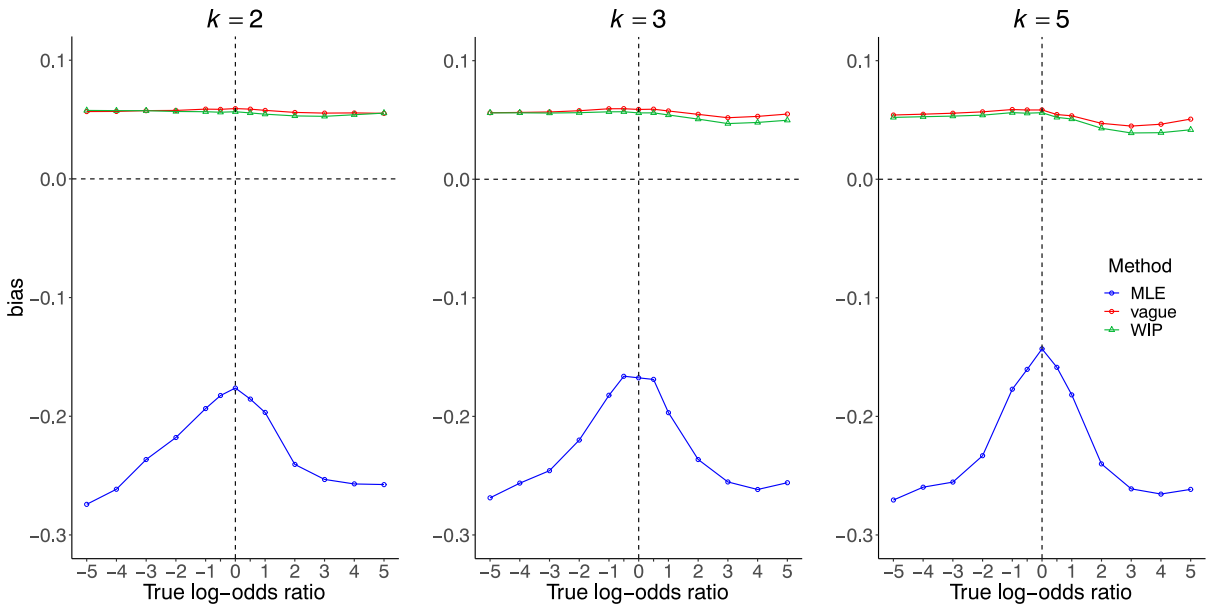

FIGURE 5 The bias for the heterogeneity parameter T obtained by three methods (maximum likelihood estimate [MLE], Vague, and weakly informative prior [WIP]) is shown. True heterogeneity standard deviation is assumed to be T=0.28 [Colour figure can be viewed at wileyonlinelibrary.com]

use the bias ($\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}-\theta)$) based on the MLE, the MH estimator, and posterior medians. The direction of the bias is also important, since depending on the nature of the outcome (safety or efficacy), a positive or a negative bias may be considered *conservative* . Moreover, the coverage probability and the mean length of interval estimates for θ are reported. The coverage probability of 95% for interval estimates and shorter interval estimates are desirable.

The bias of posterior medians from the Vague, the WIP and the BBM, and for MLEs from the MLE and for the MH estimator from the MH across scenarios is displayed in the first row of Figure 4. Note that failed runs were excluded from the calculation of performance measures that is relevant only for MLE and MH. The MLE shows unacceptably high bias for the scenarios with $\theta \leq 0$, corresponding to the scenarios in which the fraction of zero studies is also very high. On the other hand, the MH estimator clearly outperforms the MLE and exhibit bias very close to the WIP. The WIP displays somewhat positive bias whereas the Vague shows negative bias for the scenarios with $\theta \leq 0$. This behavior of WIP is expected, since the WIP shrinks the posterior towards zero. For safety analyses, a positive bias commonly means a more conservative behavior and may hence be considered less harmful than a negative bias. It is important to note that the results of the bias behave similar to the fraction of zero studies and the fraction of estimation failure of the MLE, meaning that the bias is higher in scenarios with more sparse data. Since the Vague approach uses a vague prior on θ, one might expect a somewhat similar behavior of bias from the Vague and the MLE approaches. However, the fact that the Vague approach includes a WIP for τ and that estimation is based on integration rather than maximization may be explanations of the better performance of the Vague method in comparison with the MLE. The WIP and the MH outperforms the BBM in terms of bias across all scenarios. Performance in terms of bias is improving for all methods when the number of studies k is increasing. For Figures 3 and 4, the curves are not symmetric around zero. This asymmetry is due to the fact that while the true treatment effect (log-OR) is varied between -5 and $+5$, the true baseline risk (probability) is drawn uniformly between 0.005 and 0.05 in the simulations.

Figure 4 also shows coverage probabilities and log mean lengths for 95% HDI obtained by the Vague, the WIP, the BBM, and for 95% Wald confidence intervals (CIs) obtained by the MLE and the MH. The CI and HDI obtained by the MH and the BBM show unacceptably low coverage especially for $\theta < -2$. However, the undercoverage of the BBM and somewhat relative poor performance in terms of bias may stem from the fact that data are generated under the BNHM. Also, the CI obtained by the MLE displays low coverage especially for $k = 5$. We will revisit the coverage of the MLE in the discussion. The WIP method shows higher coverage than nominal level across all different true treatment effects except for $\theta = -5$. On the other hand, the HDI obtained by WIP are shorter in comparison with HDI obtained by the Vague and CI obtained by the MLE approaches.

Lastly, the bias for the heterogeneity parameter τ obtained by three methods (the MLE, the Vague, and the WIP) are demonstrated in Figure 5. For Bayesian methods, posterior medians are used as the point estimates. Recall that the prior used for τ both in the Vague and the

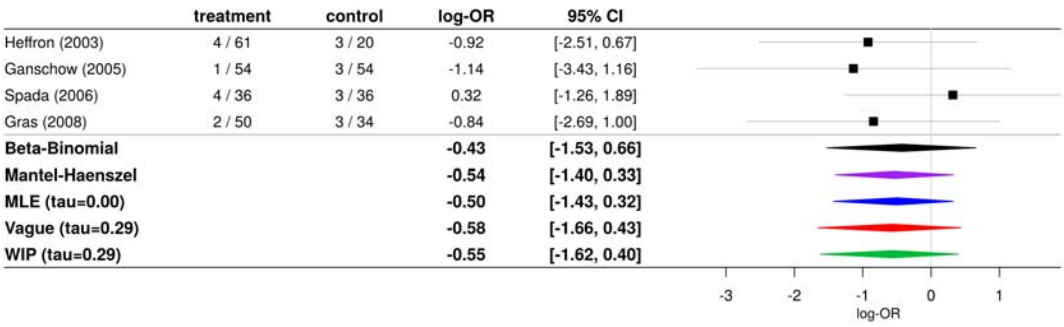

FIGURE 6 The motivating pediatric transplantation application when the outcome is death: Top panel displays the observed log-odds ratios (computed using a continuity correction in case of zero counts). The bottom panel shows mean treatment effect estimates of θ obtained by beta-binomial model (BBM), Mantel-Haenszel (MH), maximum likelihood estimate (MLE), Vague, and weakly informative prior (WIP). Heterogeneity parameter estimates τ are also given on the left [Colour figure can be viewed at wileyonlinelibrary.com]

WIP is weakly informative ($\mathcal{HN}(0.5)$). The MLE underestimates the true heterogeneity, whereas the Vague and the WIP methods slightly overestimate it. The Vague and the WIP produce very similar bias. These observations are in alignment with the conclusions made by Friede et al.[24]

6 | EXAMPLE REVISITED

Returning to the dataset described in Section 2, we consider the data on death and PTLD outcomes shown in Table 1. The observed log-odds ratios are displayed in Figures 6 and 7. To be able to visualize the observed log-odds ratios when there is a single-zero or double-zero study, a continuity correction of 0.5 is added to all cells of the single-zero or double-zero study's contingency table. The wide CI for observed log-odds ratios reflect the rather small sample sizes in the datasets. Furthermore, the variability in the point estimates may be reflected upon to assess the degree of heterogeneity between trial estimates.

We analyze the datasets using the five methods investigated in the simulation studies, namely, the Vague, WIP, MLE, MH, and BBM approaches. The code to implement the MLE and the MH are given in Appendix B. Recall that the only difference between Vague and WIP is the prior used for the treatment effect parameter θ in the model, namely, $\mathcal{N}(0, 100^2)$ for the former and $\mathcal{N}(0, 2.82^2)$ for the latter. WIP can be implemented in a routine data analysis using our `MetaStan` package as follows:

The argument `delta` corresponds to δ from (5) and thus is used to calculate the WIP for θ. Alternatively, one can directly specify the prior parameters for θ, in our case, equivalently, we can have `theta_prior = c(0, 2.82)`. The Vague method is simply implemented by omitting the argument `delta` and specifying `theta_prior = c(0, 100)`. The BBM is also implemented in `MetaStan`, and the required syntax is shown in Appendix C. To check MCMC convergence, we use the Gelman-Rubin statistics and traceplots. For the WIP approach, the corresponding traceplots are shown in Figures A1 and A2 for death and PTLD outcomes, respectively. There was no divergence reported for both datasets. The MLE fit and the MH estimation for the dataset where death is the outcome does not cause any warning from `lme4` and `metafor`, respectively. For the PTLD outcome, `lme4` gives a warning suggesting that the estimates may not be reliable. Nevertheless, it produces the MLE estimate and CI for treatment effect parameter, and we report them. For PTLD outcomes, when computing the MH estimator, `metafor` gives a warning due to double-zero studies (double-zero studies are removed from the analysis by default) but still returns an estimate. Note that both MLE and MH ignore the double-zero study (Schuller et al[30]); hence, the analyses are based on two studies only.

The results for the death and PTLD outcomes from the five methods are shown in Figures 6 and 7, respectively. For MLE and MH, the estimates and 95% CI are given. For

```
bnhm.wip.CrinsPTLD.stan  = meta_stan(ntrt = exp.total,
                                     nctrl = cont.total,
                                     rtrt = exp.PTLD.events,
                                     rctrl = cont.PTLD.event,
                                     data = dat.Crins2014.PTLD,
                                     tau_prior_dist = "half-normal",
                                     tau_prior = 0.5,
                                     delta = 250)
```

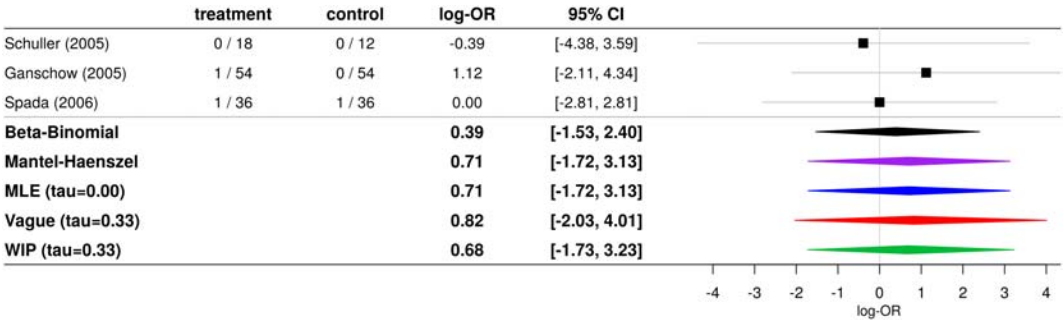

FIGURE 7 The motivating pediatric transplantation application when the outcome is posttransplant lymphoproliferative disease (PTLD): Top panel displays the observed log-odds ratios (computed using a continuity correction in case of zero counts). The bottom panel shows mean treatment effect estimates of θ obtained by beta-binomial model (BBM), Mantel-Haenszel (MH), maximum likelihood estimate (MLE), Vague, and weakly informative prior (WIP). Heterogeneity parameter estimates $\top$ are also given on the left [Colour figure can be viewed at wileyonlinelibrary.com]

Vague, WIP, and BBM, posterior medians and 95% HDI are shown. Both for PTLD and death outcomes, apart from the BBM, the point estimates of θ from the four methods look quite similar. The differing behavior of the BBM was also observed in the simulations. The PTLD data are similar to the scenarios when the number of studies is three, and the true treatment effect is in the range from 0 to 1. Negative bias obtained by the BBM can be seen in Figure 5 (in the corresponding scenario). The death data are similar to the scenarios when number of studies is five (since it is not highly sparse), and true treatment effect is in the range from -1 to 0. Here, positive bias obtained by the BBM can be seen in Figure 5 (in the corresponding scenario). Furthermore, the point estimates obtained by the WIP and the MH are very close as in the simulations. MLE gives shorter interval estimates compared with Bayesian alternatives, this is (partly) because τ was estimated as 0. In the original paper, Crins et al[26] fitted a normal-normal hierarchical model using REML,[27] and the risk ratio was used as the measure of the treatment effect. They concluded that treatment IL-2RA failed to show statistically significant result for reducing death. We obtained similar point estimates with somewhat wider interval estimates to Crins et al,[26] specifically their risk ratio estimate was 0.61 (CI, 0.27-1.37), and we obtained the odds ratio estimate 0.58(HDI 0.20-1.49) using the WIP method. Concerning the PTLD, the risk ratio was estimated as 1.60 (CI, 0.20-12.67) by Crins et al,[26] the odds ratio is estimated 1.98 (HDI 0.18-25.18) using the WIP method. The wider interval estimates obtained by WIP may stem from the fact that the uncertainty in τ is taken into account.

The estimates of the between-trial heterogeneity $\hat{\tau}$ are also included in the figure, which are only available for the Vague, WIP, and MLE. Considering death as outcomes, the heterogeneity parameter τ is estimated 0.29, 0.29, and 0.00 using WIP, Vague, and MLE, respectively. Similarly for the PTLD outcomes, for $\hat{\tau}$, we obtained 0.33, 0.33, and 0.00 using WIP, Vague, and MLE, respectively. The heterogeneity parameter of the BBM ρ is estimated as 0.34 and 0.03 for PTLD and death outcomes, respectively. Moreover, Crins et al[26] concluded that there is no evidence for heterogeneity between trials using using Cochrane's Q test for both death and PTLD outcomes. Since the prior used for τ is the same for WIP and Vague, similar heterogeneity estimates are expected. The similar τ estimates by WIP and Vague were also observed in the simulations (Figure 5). On the other hand, the MLE estimate ($\hat{\tau} = 0.00$) is most probably underestimating the actual amount of heterogeneity. The underestimation of τ by MLE and slightly lower bias of the WIP compared with the Vague was observed in the simulations (Figure 5).

7 | CONCLUSIONS AND DISCUSSION

An assumption of the homogeneity is often considered unrealistic for meta-analyses in biomedical sciences; hence, random-effects meta-analysis models are suggested.[6] Furthermore, as can be seen in the CDSR, a substantial fraction of published meta-analyses is based on few studies only. On the other hand, fitting a random-effects models based on only few studies often poses problems for inference, as certain asymptotics cannot be relied upon.[47] Additional issues arise for binary outcomes when only few or no events are observed in some of the studies or study arms. To deal with such data sparsity in the meta-analysis, we have proposed the use of WIPs for the treatment effect parameter θ in a BNHM. We demonstrated how a normal WIP for θ can be derived by considering an a priori interval for the treatment effect on a log-odds ratio scale. Also, the empirical evidence obtained from 37 773 meta-analyses with binomial outcomes from the CDSR supports the proposed WIP. In simulation studies, the suggested method displays lower bias for θ and

substantially shorter interval estimates for θ with somewhat higher coverage than nominal level in comparison to alternative methods.

The use of a Bayesian approach exhibits analogy of some degree to the use of continuity corrections. While continuity corrections might to some extent be perceived as ad hoc makeshift fixes, they have also quite doubtlessly proven very useful in practice. A Bayesian approach tackles the problem from a very different angle, but it is not so surprising that the resulting procedure again exhibits some similarity to continuity corrections. The relation to current common practice may in fact also be seen as somewhat comforting. Use of an (informative) prior within a Bayesian analysis on the other hand is not a desperate measure; it is rather an integral part of a coherent model specification that may also be subjected to checks of plausibility and operating characteristics; this is what we have tried to demonstrate in the present paper.

The simulation results displayed in Figure 4 are somewhat in contrast to the results given by Friede et al,[24] who observed lower coverage than nominal level of MLE methods in a similar setting, but not based on rare events. We also investigated a scenario closer to their setup by considering higher baseline risks between 0.05 and 0.20. The results are shown in Figure D1, and indeed, here, the MLE method exhibits lower coverage than nominal level, as reported by Chung et al[22] and Friede et al.[24] The high bias and too wide interval estimates obtained by the Vague and the MLE are still present, but not as high as in the results of the simulations in which true baseline risks are lower.

Jackson et al[11] investigated seven random-effects meta-analysis models including the BNHM which we consider in this paper (*model 4* in Jackson et al[11]) and another parametrization of the BNHM (*model 2* in Jackson et al[11]). The only difference in the specification between the two models is that in their Model 2, the treatment indicator x_{ik} of (3.1) is +1 for the experimental arm, and 0 for the control arm. Note that commonly used network meta-analysis models, for example,[48] are generalizations of Model 2 in Jackson et al.[11] As reported by Jackson et al,[11] we also observe the underestimation of the heterogeneity parameter τ and hence decided to only consider their model 4. On the other hand, it is important to note that the usage of a WIP for θ also improves the performance in model 2, as we have seen for the model 4.

This investigation has some limitations. One crucial limitation is that we only considered the BNHM as a data-generating process in our simulation study. Hence, we did not investigate the robustness of the BNHM under model misspecification. Also, the design of the simulation study constitutes a model misspecification problem for the MH method, which is a fixed-effect model, and for the BBM, which assumes a different underlying data-generating process. Moreover, we did not consider other parametrizations of the BNHM as described, eg, in Jackson et al.[11] Lastly, one may find it too restrictive to have a normal prior for θ as we have in our proposed model, it may be worth exploring alternatives like Cauchy or log-F distributions[17,20] for penalization.

The proposed approach is not restricted to the BNHM; similar approaches may analogously be defined in other models, eg, a Poisson-normal hierarchical model. However, a crucial point is that the treatment effect parameter is explicitly parameterized in the model, so that it can directly be *penalized* via the prior specification. Hence, so-called contrast-based models[25] (in which relative treatment effects are assumed to be exchangeable across trials) are suitable for this purpose unlike arm-based models. Note that this is also related to the inclusion of baseline risks as fixed effects with vague priors. This was on purpose as we consider this closest to the idea of stratifying the analyses by study, a common feature of meta-analyses regardless of fixed or random-effects. Furthermore, the contrast-based models such as the BNHM preserve the randomization, in contrast to the arm-based models as explained in Dias and Ades.[25]

The BNHM can be extended to a network meta-analysis model,[49] which is desirable if there are multiple treatments, and/or multiarm trials in the dataset. Even if the dataset in a network meta-analysis consists of many studies overall, some of the treatment effects may still be informed by few studies only. Thus, the use of WIPs for treatment effect parameters in the context of network meta-analysis with rare events can be very helpful. Different distributions as WIP for θ, different parametrizations of BNHM, or different data models can be implemented in **Stan** or MCMC methods in general. Although, currently, our package `MetaStan` is restricted to use a BNHM and BBM for pairwise meta-analysis, we will consider to extend it to conduct meta-analysis and network meta-analysis with flexible data model and prior options in the future.

ACKNOWLEDGMENT

We thank Leonhard Held who contributed valuable comments and pointed us to several important references.

HIGHLIGHTS

What is already known: Standard random-effects meta-analysis methods are not suitable for meta-analysis of few studies with rare events.

What is new: To deal with data sparsity present in the random-effects meta-analysis of few studies with rare events, we suggest the use of weakly informative priors as penalization for the treatment effect parameter.

Potential impact for RSM readers outside the authors' field: To make it more accessible to meta-analysts, a publicly available **R** package, `MetaStan`, is developed for fitting Bayesian meta-analysis models using weakly informative priors.

CONFLICT OF INTEREST

The author reported no conflict of interest.

ORCID

Burak Kürsad Günhan https://orcid.org/0000-0002-7454-8680
Christian Röver https://orcid.org/0000-0002-6911-698X
Tim Friede https://orcid.org/0000-0001-5347-7441

REFERENCES

1. Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester: Wiley; 2008.
2. Friedrich JO, Adhikari NKJ, Beyene J. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Med Res Methodol*. 2007;7(1):5. https://doi.org/10.1186/1471-2288-7-5
3. Kuss O. Statistical methods for meta-analyses including information from studies without any events–add nothing to nothing and succeed nevertheless. *Stat Med*. 2015;34(7):1097-1116.
4. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis*. 1985;27(5):335-371.
5. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst Monogr*. 1959;22(4):719-748. https://doi.org/10.1093/jnci/22.4.719
6. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc*. 2009;172(1):137-159. https://doi.org/10.1111/j.1467-985X.2008.00552.x
7. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med*. 2015;34(6):984-998.
8. Kontopantelis E, Springate DA, Reeves D. A re-analysis of the cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *PLoS One*. 2013;8(7):1-14.
9. Hedges LV, Olkin I. Random effects models for effect sizes. *Statistical methods for meta-analysis*; 1985.
10. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med*. 2007;26(1):53-77. https://doi.org/10.1002/sim.2528
11. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat Med*. 2018;37(7):1059-1085.
12. Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med*. 2010;29(29):3046-3067.
13. Böhning D, Mylona K, Kimber A. Meta-analysis of clinical trials with rare events. *Biom J*. 2015;57(4):633-648. https://doi.org/10.1002/bimj.201400184
14. Cai T, Parast L, Ryan L. Meta-analysis for rare events. *Stat Med*. 2010;29(20):2078-2089. https://doi.org/10.1002/sim.3964
15. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med*. 1995;14(24):2685-2699. https://doi.org/10.1002/sim.4780142408
16. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 1984;71(1):1-10.
17. Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat Med*. 2015;34(23):3133-3143.
18. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38.
19. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med*. 2002;21(16):2409-2419.
20. Gelman A, Jakulin A, Pittau MG, Su Y. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat*. 2008;2(4):1360-1383.
21. Davey J, Turner RM, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Med Res Methodol*. 2011;11(1):160.
22. Chung Y, Rabe-Hesketh S, Choi IH. Avoiding zero between-study variance estimates in random-effects meta-analysis. *Stat Med*. 2013;32(23):4071-4089.
23. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Anal*. 2006;1(3):515-534. https://doi.org/10.1214/06-BA117A
24. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. *Res Synth Methods*. 2017;8(1):79-91. https://doi.org/10.1002/jrsm.1217
25. Dias S, Ades AE. Absolute or relative effects? Arm-based synthesis of trial data. *Res Synth Methods*. 2016;7(1):23-28. https://doi.org/10.1002/jrsm.1184
26. Crins ND, Röver C, Goralczyk AD, Friede T. Interleukin-2 receptor antagonists for pediatric liver transplant recipients: a systematic review and meta-analysis of controlled studies. *Pediatr Transplant*. 2014;18(8):839-850. https://doi.org/10.1111/petr.12362
27. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48. http://www.jstatsoft.org/v36/i03/
28. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327:557-560.
29. Heffron TG, Pillen T, Smallwood GA, Welch D, Oakley B, Romero R. Pediatric liver transplantation with daclizumab induction therapy. *Transplant*. 2003;75:2040-2043.
30. Schuller S, Wiederkehr JC, Coelho-Lemos IM, Avilla SG, Schultz C. Daclizumab induction therapy associated with tacrolimus-mmf has better outcome compared with tacrolimus-MMF alone in pediatric living donor

liver transplantation. *Transplant Proc*. 2005;37(2): 1151-1152.
31. Ganschow R, Grabhorn E, Schulz A, Hugo AV, Rogiers X, Burdelski M. Long-term results of basiliximab induction immunosuppression in pediatric liver transplant recipients. *Pediatr Transplant*. 2005;9(6):741-745.
32. Spada M, Petz W, Bertani A, et al. Randomized trial of basiliximab induction versus steroid therapy in pediatric liver allograft recipients under tacrolimus immunosuppression. *Am J Transplant*. 2006;6(8):1913-1921.
33. Gras JM, Gerkens S, Beguin C, et al. Steroid-free, tacrolimus-basiliximab immunosuppression in pediatric liver transplantation: clinical and pharmacoeconomic study in 50 children. *Liver Transpl*. 2008;14(4):469-477.
34. Spiegelhalter DJ, Abrams KR, Myles JP. Prior distributions. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. West Sussex: CRC Press; 2004.
35. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol*. 2006;35(3):765-775.
36. Kass RE, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *J Am Stat Assoc*. 1995;90(431):928-934.
37. Röver C. Bayesian random-effects meta-analysis using the bayesmeta R package. arxiv.org e-print archive. https://arxiv.org/abs/1711.08683 updated november 23, 2017. accessed August 30, 2018.
38. The cochrane. cochrane database of systematic reviews. https://www.cochranelibrary.com accessed april 15, 2018.
39. Springate D. Cochrane_scraper v1.1.0. https://doi.org/10.5281/zenodo.10782 updated july, 2014. accessed april, 2018.
40. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the cochrane Database of Systematic Reviews. *Int J Epidemiol*. 2012;41(3):818-827. https://doi.org/10.1093/ije/dys041
41. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1-48.
42. Stan Development Team. Stan modeling language user's guide and reference manual, version 2.17.0.; 2017.
43. Betancourt M. A conceptual introduction to Hamiltonian Monte Carlo. arxiv.org e-print archive. https://arxiv.org/abs/1701.02434. updated july 16, 2018. accessed august 24, 2018.
44. Betancourt M, Girolami M. Hamiltonian Monte Carlo for hierarchical models. *Current Trends in Bayesian Methodology With Applications*; 2015.
45. Meredith M, kruschke J. HDInterval: Highest (posterior) density intervals, R package version 0.2.0.https://CRAN.R-project.org/package=HDInterval; 2018.
46. R core team. *R: A language and environment for statistical computing*; R foundation for statistical computing. https://www.R-project.org/ Vienna, Austria. 2018.
47. Jackson D, White IR. When should meta-analysis avoid making hidden normality assumptions?. *Biom J*. 2018;60(6):1040-1058. https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201800071
48. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical support document 2: A generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. http://www.nicedsu.org.uk last updated September 2016; 2011.
49. Günhan BK, Friede T, Held L. A design-by-treatment interaction model for network meta-analysis and meta-regression with integrated nested laplace approximations. *Res Synth Methods*. 2018;9(2):179-194.

How to cite this article: Günhan BK, Röver C, Friede T. Random-effects meta-analysis of few studies involving rare events. *Res Syn Meth*. 2020;11: 74–90. https://doi.org/10.1002/jrsm.1370

APPENDIX A: HOW TO USE THE METASTAN R PACKAGE

The stable version of **MetaStan** is available on CRAN (https://CRAN.R-project.org/package=MetaStan) and can be installed as follows:

```
install.packages("MetaStan")
```

The example described in the text (Crins dataset) is available in the package, and it can be loaded as follows:

```
library("MetaStan")
data("dat.Crins2014")
## Subset of dataset where PTLD outcomes available
dat.Crins2014.PTLD = subset(dat.Crins2014, is.finite(exp.PTLD.events))
```

Additional information can be obtained by typing `?dat.Crins2014` (for any dataset and function in the package).

`meta_stan` is the main fitting function of this package. The main computations are executed via the **rstan** package's `sampling` function. We can fit the binomial-normal hierarchical using a WIP for treatment effect as follows:

```
bnhm.wip.CrinsPTLD.stan  = meta_stan(ntrt = exp.total,
                                     nctrl = cont.total,
                                     rtrt = exp.PTLD.events,
                                     rctrl = cont.PTLD.event,
                                     data = dat.Crins2014.PTLD,
                                     tau_prior_dist = "half-normal",
                                     tau_prior = 0.5,
                                     delta = 250)
                                     chains = 4,
                                     iter = 2000,
                                     warmup = 1000)
```

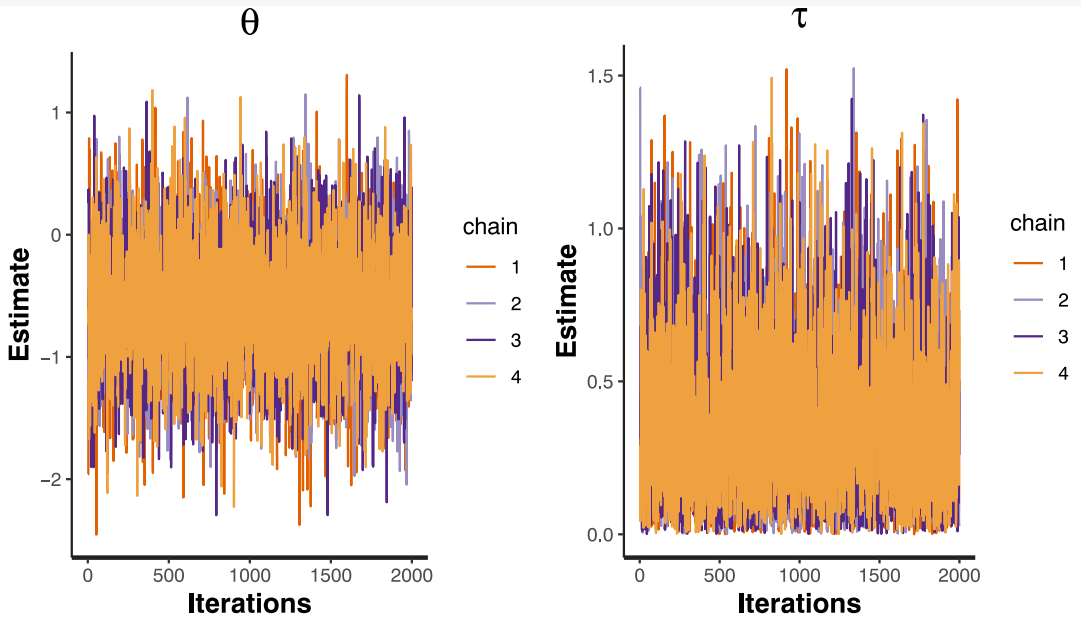

FIGURE A1 Traceplots for the estimated parameters θ and τ including burn-in for death outcomes [Colour figure can be viewed at wileyonlinelibrary.com]

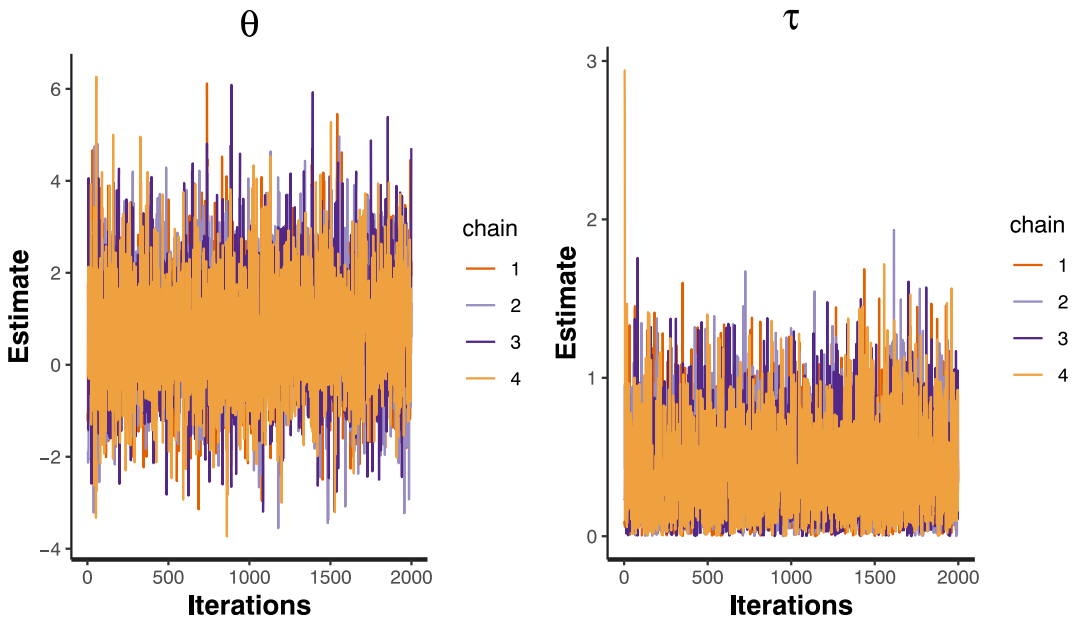

FIGURE A2 Traceplots for the estimated parameters θ and τ including burn-in for posttransplant lymphoproliferative disease (PTLD) outcomes [Colour figure can be viewed at wileyonlinelibrary.com]

Convergence diagnostics and the results can be very conveniently obtained using the **shinystan** package as follows:

```
library("shinystan")
## Firstly convert "stan" object to a "shinystan" object
bnhm.wip.CrinsPTLD.shinystan = as.shinystan(bnhm.wip.CrinsPTLD.stan$fit)
launch_shinystan(bnhm.wip.CrinsPTLD.shinystan)
```

Traceplots for the estimated parameters θ and τ including burn-in are shown in Figures A1 and A2 for death and PTLD outcomes, respectively.

Lastly, the posterior summary statistics can be obtained using the following command:

```
bnhm.wip.CrinsPTLD.stan$fit_sum
```

APPENDIX B: R CODE TO IMPLEMENT BNHM USING THE MLE AND THE MH METHODS

Firstly, the BNHM using the MLE:

```
library("lme4")
## Firstly convert dataset to a long format
## using MetaStan::convert_data_arm function
dat.Crins2014.PTLD.long = convert_data_arm(dat.Crins2014.PTLD$exp.total,
                                           dat.Crins2014.PTLD$cont.total,
                                           dat.Crins2014.PTLD$exp.PTLD.events,
                                           dat.Crins2014.PTLD$cont.PTLD.events)

glmer(cbind(r, sampleSize - r) ~ factor(mu) + factor(theta) + (theta12 - 1|mu),
      data = dat.Crins2014.PTLD.long, family = binomial(link = "logit"), nAGQ = 7)
```

Secondly, the MH method:

```
library("metafor")
rma.mh(n1i = exp.total, n2i = cont.total,
       ai = exp.PTLD.events, ci = cont.PTLD.event,
       data = dat.Crins2014.PTLD, measure = "OR")
```

APPENDIX C: R CODE TO IMPLEMENT THE BBM METHOD

```
bbm.CrinsPTLD.stan  = meta_stan(ntrt = exp.total,
                                nctrl = cont.total,
                                rtrt = exp.PTLD.events,
                                rctrl = cont.PTLD.event,
                                data = dat.Crins2014.PTLD,
                                model = "Beta-binomial")
```

APPENDIX D: ADDITIONAL SIMULATION RESULTS

We also conducted simulations using the same settings as described in Section 5 under BNHM, but using higher baseline risk probabilities, specifically, baseline risks (μ_i) are uniformly taken between 0.05 and 0.2. Results are illustrated in Figure D1 (analogous to Figure 5).

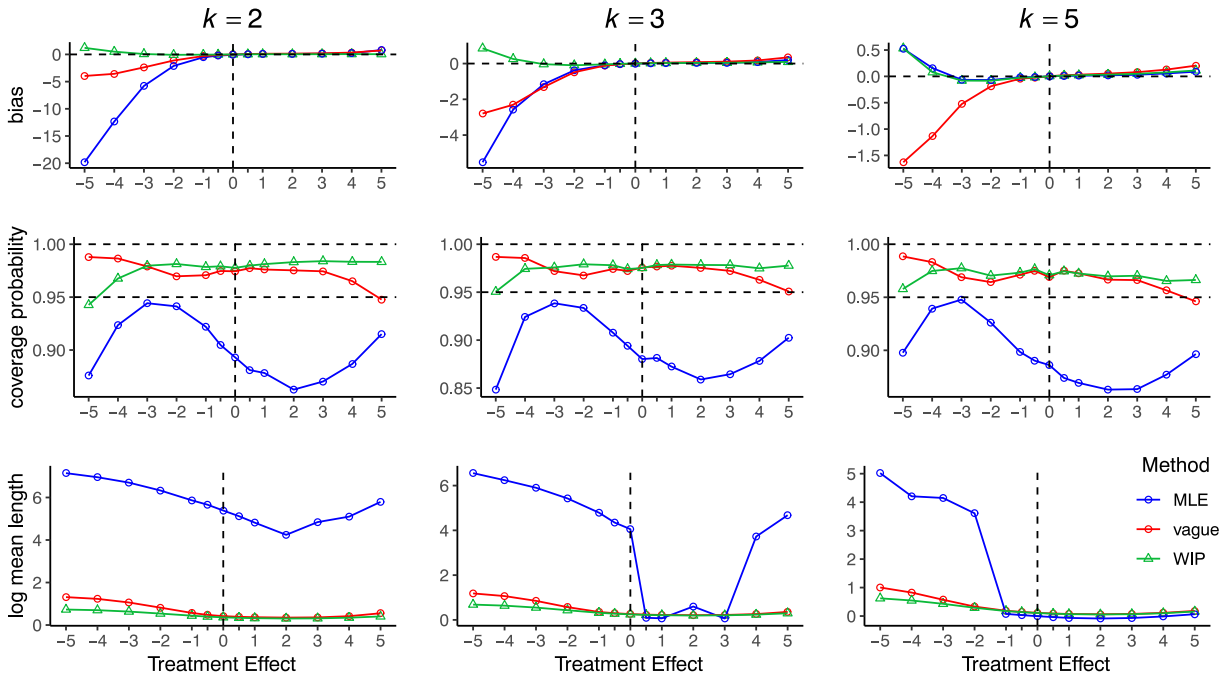

FIGURE D1 Simulations with high baseline risks: The bias for the mean treatment effect θ, coverage probabilities, and log mean length of the interval estimates for θ obtained by three methods (maximum likelihood estimate [MLE], Vague, and weakly informative prior [WIP]) are shown [Colour figure can be viewed at wileyonlinelibrary.com]