Received: 30 April 2021 | Revised: 28 July 2022 | Accepted: 3 October 2022

DOI: 10.1002/sim.9593

RESEARCH ARTICLE

Statistics in Medicine WILEY

Dose finding studies for therapies with late-onset toxicities: A comparison study of designs

Helen Barnett[1,2] | **Oliver Boix**[3] | **Dimitris Kontos**[4] | **Thomas Jaki**[1,5]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

[2]Learning Development, Lancaster University, Lancaster, UK

[3]Bayer AG, Leverkusen, Germany

[4]ClinBAY, Limassol, Cyprus

[5]Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany

Correspondence
Helen Barnett, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK.
Email: helenyvetteb@gmail.com

Funding information
Medical Research Council, Grant/Award Number: MC_UU_00002/14; National Institute for Health Research, Grant/Award Number: NIHR-SRF-2015-08-001; NIHR Cambridge Biomedical Research Centre, Grant/Award Number: BRC-1215-20014

An objective of phase I dose-finding trials is to find the maximum tolerated dose; the dose with a particular risk of toxicity. Frequently, this risk is assessed across the first cycle of therapy. However, in oncology, a course of treatment frequently consists of multiple cycles of therapy. In many cases, the overall risk of toxicity for a given treatment is not fully encapsulated by observations from the first cycle, and hence it is advantageous to include toxicity outcomes from later cycles in phase I trials. Extending the follow up period in a trial naturally extends the total length of the trial which is undesirable. We present a comparison of eight methods that incorporate late onset toxicities while not extensively extending the trial length. We conduct simulation studies over a number of scenarios and in two settings; the first setting with minimal stopping rules and the second setting with a full set of standard stopping rules expected in such a dose finding study. We find that the model-based approaches in general outperform the model-assisted approaches, with an interval censored approach and a modified version of the time-to-event continual reassessment method giving the most promising overall performance in terms of correct selections and trial length. Further recommendations are made for the implementation of such methods.

KEYWORDS
dose-finding, late-onset toxicities, model-assisted, model-based, phase I trials

1 | INTRODUCTION

In phase I dose finding studies, the aim is to find the maximum tolerated dose (MTD) to recommend for phase II, defined as the highest dose giving an acceptable level of toxicity.[1] An acceptable level of toxicity is in general equated to a certain probability of occurrence of a dose limiting toxicity event (DLT), which in oncology is most often defined as a grade 3 or higher toxicity by the grading scale of the National Cancer Institute.[2] This probability is usually set to 20%-30% within the general treatment population, with the actual choice of target probability depending on the indication, available treatment option and expected benefit within the target population.

However, this probability refers to the risk of DLT within the follow-up period of the trial, which is typically only one cycle of treatment. In many cancer treatments, the full course of treatment actually consists of multiple cycles of therapy, given sequentially. It is therefore important to consider multiple cycles of therapy in the dose-finding trial. In a review

of 445 patients in 36 phase I trials by Postel-Vinay et al,[3] it was found that 57% of grade 3 or 4 toxicities occurred after the first cycle of treatment, and that for 50% of patients, their worst grade of toxicity was observed after the first cycle. This is a clear indication that later cycles are important to include in phase I dose-finding trials, since their omission may lead to missing large amounts of information on toxicity risk of the investigated doses and hence the recommendation of sub-optimal doses. However, by increasing the follow up period, one also greatly increases the trial duration if the entire follow up period must be completed for the previous cohort of patients before the dose for the next cohort can be assigned. Such an increase in trial duration is obviously undesirable, as the focus in such trials is in efficient decision making.

A small number of statistical methods for the design of trials in such a setting have been proposed, with varying approaches to include the later onset toxicities without substantially increasing the trial duration. For example, the Rolling 6 design,[4] a rule based approach that is an extension of the 3+3,[5] uses a set of rules based on the number of DLTs observed and the number of patients who have completed and are yet to complete their full follow up period. Other (model-assisted) designs,[6-8] follow similar sets of rules, with the addition that escalation is aided by a simple model for the probability of toxicity at each dose level under the assumption of monotonicity, that a higher dose is associated with a higher probability of toxicity. There are also a limited selection of model based designs that account for later onset toxicities.[9-12] These designs vary in their approach to accounting for the occurrence of DLTs in the different treatment cycles. The time-to-event approach[9] models the DLT occurrence using a time-to-event variable defined on the entire follow up period, and not necessarily breaking down this period into cycles. The other model-based approaches do break the follow up period into the respective cycles, with the interval censored approach of Sinclair and Whitehead[10] modeling the probability of DLT in each cycle, conditional on the lack of DLT in all previous cycles, whereas the method by Doussau et al[11] fits a proportional-odds mixed model to data from the different cycles. The approach by Yin et al[12] fits a linear mixed effects model including a cycle effect to data of total toxicity profile, including grades and type of toxicity.

As expected with novel statistical methodology, each approach is praised by the respective authors for its advantages over another given method in any particular setting. However, such settings usually fit well to the approach suggested, and although some exploration of settings that violate assumptions is often undertaken, it would be of great aid to have a comparison of the leading methods in settings that are both realistic and not adhering to the assumptions of the approaches. Although different approaches use different levels of information in their analyses, it does not necessarily mean that a higher level of information used leads to a higher level of accuracy, since higher levels of information often accompany stronger assumptions in these methods. More complex methods may also struggle with the small sample sizes in such studies. Therefore in this work, we undertake a simulation study to compare the most prominent methods for dose-finding studies incorporating late-onset toxicities, in order to evaluate the strengths and weaknesses of each of the methods, and their applicability to phase I dose finding studies. We have chosen this range to incorporate different levels of complexity and to include well-known and already applied designs. A strength of such a wide range of designs is that they have such different approaches. In this comparison, we use modified versions of the methods, to improve their applicability in this setting and to ensure their comparability.

This article has the following structure. In Section 2, the eight approaches for comparison are outlined, with notations introduced and assumptions of each method highlighted. In Section 3, we introduce the setting for the simulation study, and in Section 4, we describe the procedure we use to choose the values for the hyper-parameters of the prior distributions in order for a fair comparison across methods. In Section 5, we present the results of the simulation study, before concluding with a discussion in Section 6.

2 | METHODS

In this section, we outline the eight methods that are implemented in the comparison study. The purpose is to give an overview of each method, with key details on the different models used in the dose escalation. Further and more in-depth descriptions can be found in the relevant referenced literature. We highlight any modifications made to the original proposals.

In each method, the following notation is consistent throughout. Consider a dose-finding study where J dose levels labeled d_j for $j = 1, \dots, J$ are investigated. Each patient i is followed up for S cycles of therapy indexed $s = 1, \dots, S$. A new cohort of patients is admitted at the beginning of each cycle, so that partial information is available for the previous $S - 1$ cohorts when assigning the dose for the current cohort. For example, for a follow up period of 3 cycles, partial information

is available for the previous 2 cohorts; information is available for only the first cycle of therapy of the previous cohort, and information on the first two cycles for the second previous cohort. Dose assignment is based on some target, τ, defined as the probability of observing a DLT in the entire follow-up of S cycles.

The general process of a dose-finding study is as follows. Pretrial, for a model-based method, a dose response model is chosen and prior distributions assigned to the parameters. For a model-assisted method, prior distributions are assigned to any relevant parameters that guide escalation, and a decision table based on these and all possible outcomes is calculated. The first cohort of patients is assigned to the lowest dose. After one cycle of treatment, the observed responses from these patients are used to decide which dose to assign to the next cohort of patients to, or to stop the trial. For a model-based method the posterior distribution is updated from which the next best dose is derived and for a model-assisted method the result is looked up in the decision table. This same process is repeated after each cycle of treatment, until the trial is stopped for a prespecified reason, and either a dose is recommended as the MTD or all doses are deemed unsafe in which case no MTD is recommended.

2.1 | Time-to-event continual reassessment method approach

The first method we review is perhaps the most well known, the time-to-event continual reassessment method approach (TITE-CRM) first proposed by Cheung and Chappell.[9] We consider two approaches under the umbrella of TITE-CRM, an approach closely mirroring the original proposed methodology by Cheung and Chappell[9] (1-parameter TITE-CRM) and a modification to include the actual dose values instead of standardized doses (2-parameter TITE-CRM).

2.1.1 | 1-Parameter TITE-CRM

Here, a weighted dose response model is used:

$$G(d, w, \beta) = wF(d, \beta),$$

where $0 \leq w \leq 1$ is a function of time-to-event of a patient response, F is the assumed dose-response model, d is the scaled dose, and β is the parameter to be estimated. The scaled doses are interpreted as the prior belief of the probability of toxicity on that dose, these are used as opposed to real values of the dosages. The dose-response model suggested is $F(d, \beta) = d^{\exp(\beta)}$ and a Normal prior distribution is elicited on β ($\sim N(0, \sigma^2)$), with the posterior distribution updated after each cycle using likelihood

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} G(d_{[i]}, w_{i,n}, \beta)^{y_{i,n}^{(TC)}} \{1 - G(d_{[i]}, w_{i,n}, \beta)\}^{1-y_{i,n}^{(TC)}},$$

where n is the number of patients treated, and $y_{i,n}^{(TC)}$ is an indicator taking the value 1 if patient i has observed a DLT after information is available for at least one cycle for n patients. Dose assignment is determined by minimizing $|F(d_j, \hat{\beta}_n) - \tau|$ where $\hat{\beta}_n$ is the posterior mean of β after information is available for at least one cycle for n patients and τ is the target P(DLT) for all S cycles. We use the weights suggested by Cheung and Chappell,[9] of the simple $w_{i,n} = u_{i,n}/S$, where $u_{i,n}$ is the current number of cycles patient i has been observed for. As outlined by Cheung and Chappell,[9] if a DLT is observed then $w_{i,n} = 1$. The final dose recommendation is the dose level that minimizes $|F(d_j, \hat{\beta}_n) - \tau|$ once the follow-up for all enrolled patients has completed.

Although this method is flexible enough to allow for continuous time-to-event responses, here we discretize this variable according to the cycle the response is observed in. The TITE-CRM has an initial period where the dose is escalated one level at a time until a DLT is observed. In the original methodology, in this initial period, each patient is followed up for their entire follow up time before the next patient's dose is assigned. In our implementation, only one cycle is required for follow up before the next is assigned, in line with the rest of the trial. This is also in line with the other methods considered in this comparison.

2.1.2 | 2-Parameter TITE-CRM

In order to make use of the real doses used in the trial, we modify the TITE-CRM to include these, henceforth labeled as the TITE-CRM2. Since the above dose-response model is only valid for $0 < d < 1$ we now use the 2-parameter logistic model:

$$F(d_j, \boldsymbol{\beta}) = \frac{\exp(a_0 + a_1 d_j)}{1 + \exp(a_0 + a_1 d_j)},$$

where d_j is the real value of the dose. Here, MCMC methods must be used to update the posterior distribution for $\boldsymbol{\beta} = (a_0, a_1)$. The prior distributions are $a_0 \sim N(\mu_{a_0}, \sigma^2_{a_0})$ and $\log(a_1) \sim N(\mu_{a_1}, \sigma^2_{a_1})$. These Normal priors are in line with other dose-finding methods.[13]

2.2 | Interval censored survival approach

This approach introduced by Sinclair and Whitehead[10] uses $\pi_{j,s}$, the conditional probability of observing a DLT in cycle s for a patient on dose d_j given they did not observe a DLT in previous cycles. The prior for this method takes the form of pseudo-data, based on the approach by Whitehead and Williamson,[14] where a small number of pseudo-patient observations (allowing for non-integer observations) used to guide the escalation. Here we have n_0 pseudo-patient observations on dose d_1, with $\pi^*_1 n_0$ patients observing a DLT on the first cycle and n_0 pseudo-patient observations on dose d_J, with $\pi^*_J n_0$ patients observing a DLT on the first cycle. Pseudo-data for subsequent cycles is calculated based on a decreasing π^*_j.

With the pseudo-data prior, the posterior for θ and $\boldsymbol{\gamma}$ is updated using likelihood:

$$\mathcal{L}(\theta, \boldsymbol{\gamma}) = \prod_{j=1}^{J} \prod_{s=1}^{S} \pi_{j,s}^{r_{j,s}} (1 - \pi_{j,s})^{q_{j,s}},$$

with link function $\log(-\log(1 - \pi_{j,s})) = \gamma_s + \theta \log(d_j)$, where $r_{j,s}$ is the number of patients that experience a DLT in cycle s, $q_{j,s}$ is the number of patients who have completed s cycles without experiencing a DLT, γ_s is an intercept term for cycle s. Dose assignment is determined by maximizing gain function $1/(\tau - \hat{\rho}_j(t_S))^2$ where τ is the target $P(\text{DLT})$ for all S cycles and $\hat{\rho}_j(t_S)$ is the current estimate of $P(\text{DLT})$ for dose d_j for all S cycles. The final dose recommendation is the dose level that maximizes $1/(\tau - \hat{\rho}_j(t_S))^2$ once the follow-up for all enrolled patients has completed. This is a modification from the original proposal, which recommended any dose on the continuous scale, which we make in order for results of this design be measured by the same metric as other designs.

2.3 | Proportional odds mixed effect model approach

The third approach is the proportional odds mixed effect model approach (POMM) proposed by Doussau et al.[11] This method uses the additional information on the grades of toxicity observed in each cycle. Due to difficulties in model fitting in the original proposal, we make two modifications to improve the stability of the method. We introduce the use of prior information to align this method with the other Bayesian approaches, as well as to aid the model fitting. This prior is in the form of pseudo-data, with responses from pseudo-patients on all doses and cycles, down-weighted so as to not outweigh the observed data. We also alter the target probability of DLT to include all cycles, as opposed to per cycle. Again, this aligns with other methods but also allows for differing risks across cycles.

Toxicities are categorized by grade, the response variable for subject i in cycle s is defined as:

$$y^{(\text{POMM})}_{i,s} = \begin{cases} 1, & \text{if no toxicity or grade 1 toxicity,} \\ 2, & \text{if grade 2 toxicity, } or \\ 3, & \text{if grade 3+ toxicity.} \end{cases}$$

For the first 15 subjects, the following generalized linear model is used:

$$\text{logit}\left(P\left(Y_{i,1}^{(\text{POMM})}=3\right)\middle|\, d_j\right)=b_0+b_1 d_j,$$

so that there are no mixed effects and only the responses from cycle 1 are used. Dose assignment is determined by minimizing $|P(Y_{i,1}^{(\text{POMM})})=3|\text{data},\hat{b_0},\hat{b_1},d_j)-\tau_{s=1}|$ where $\tau_{s=1}$ is the target $P(\text{DLT})$ for cycle 1.

From subject 16 onwards, a logistic proportional odds mixed-effect regression model is used

$$\text{logit}\left(P\left(Y_{i,s}^{(\text{POMM})}\le k|d_j\right)\right)=\alpha_k-\beta_1 d_j-\beta_2 s-u_i \quad \text{for } k=1,2,$$

where $u_i\sim N(0,\sigma_0^2)$ is the individual patient effect. $\boldsymbol{\theta}=(\alpha_1,\alpha_2,\beta_1,\beta_2)$. Dose assignment is determined by minimizing $|P(Y_{i,-}^{(\text{POMM})}=3|\text{data},\hat{\boldsymbol{\theta}},d_j,u_i=0)-\tau|$ where τ is the target $P(\text{DLT})$ for the whole follow up period and $P(Y_{i,-}^{(\text{POMM})})$ is the estimated probability of DLT in the whole follow up period for a patient with $u_i=0$. If a DLT is observed in any cycle, then this is dealt with in the censored likelihood.

The final dose recommendation is the dose level that minimizes $|P(Y_{i,-}^{(\text{POMM})}=3|\text{data},\hat{\boldsymbol{\theta}},d_j,u_i=0)-\tau|$ (or $|P(Y_{i,1}^{(\text{POMM})}=3|\text{data},\hat{b_0},\hat{b_1},d_j)-\tau_{s=1}|$ if there are less than 16 patients) once the follow-up for all enrolled patients has completed.

2.4 | Total toxicity profile approach

This approach proposed by Yin et al[12] calculates a normalized total toxicity profile (nTTP) value for each patient using information on grades and types of toxicity. The flexibility of this approach allows for any number of types of toxicity, provided the specification for the method of calculating the nTTP value is decided before the trial. Here, following from Yin et al,[12] three types of toxicity are included (renal, hematological, neurological), each with grades 0 to 4. Patients can observe any combination of grades and types of toxicity within a given cycle, and the maximum grade of each type for each cycle is recorded. This is used to calculate the nTTP value using the weights specified by Yin et al.[12] A linear mixed effect model is fitted to the nTTP values for all cycles:

$$y_{i,s}^{(\text{nTTP})}=\beta_0+\beta_1 x_i+\beta_2 s+\gamma_i+\epsilon_{i,s},$$

where $y_{i,s}^{(\text{nTTP})}$ is the observed nTTP value for patient i on cycle s, x_i is the dose assigned to patient i, $\gamma_i\sim N(0,\sigma_\gamma^2)$ is the individual patient effect, and $\epsilon_{i,s}\sim N(0,\sigma_\epsilon^2)$ is measurement error. The following priors are elicited: $\beta_0\sim N(\mu_{\beta_0},\sigma_{\beta_0}^2)$, $\beta_1\sim N(\mu_{\beta_1},\sigma_{\beta_1}^2)$, $\beta_2\sim N(\mu_{\beta_2},\sigma_{\beta_2}^2)$, $\sigma_\gamma^2\sim \text{IG}(0.001,0.001)$, and $\sigma_\epsilon^2\sim \text{IG}(0.001,0.001)$ and the posterior distributions for all parameters are updated after each cycle. As in the POMM method, if a DLT is observed in any cycle, then this is dealt with in the censored likelihood. Dose assignment is determined by minimizing $|\text{nTTP}(d_j,\text{cycle }1)-\tau_{\text{nTTP}}|$ where τ_{nTTP} is the target nTTP for cycle 1 and $nTTP(d_j,\text{cycle }1)$ is the current posterior estimate of nTTP for dose d_j for cycle 1. Although the nTTP value for each cycle is included in the model, only the current posterior estimate of nTTP for cycle 1 is included in the criterion. Because of the nature of the nTTP variable, there is no natural measure across all cycles in the same way there is with a variable of probability and hence we here use the criterion using only cycle 1 as recommended by Yin et al.[12]

The final dose recommendation is the dose level that minimizes $|\text{nTTP}(d_j,\text{cycle }1)-\tau_{\text{nTTP}}|$ once the follow-up for all enrolled patients has completed.

2.5 | Time to event Bayesian optimal interval approach

The time to event Bayesian optimal interval (TITE-BOIN) model-assisted approach proposed by Yuan et al[6] is a time-to-event extension of the Bayesian optimal interval design, whereby dose escalation is guided by the target interval (τ_1,τ_2). Doses are escalated or de-escalated one dose level at a time, and this decision is based on two metrics: the escalation and de-escalation boundaries of the non-time-to-event versions (λ_d & λ_e) and the standardized total follow-up time

(STFT). A set of rules based on these two values determines whether the dose is escalated, de-escalated or remained constant. A decision table for these rules can be calculated before the trial, based on the target interval and the Beta(α, β) prior assigned to the probability of DLT at each dose. The values of α and β are originally chosen so that the prior has an effective sample size of 1 and mean $\tau/2$, with this prior the same across all doses. The target interval used by the authors is $(0.6\tau, 1.4\tau)$. The final dose recommendation is based on an isotonic regression once the follow-up for all enrolled patients has completed.

2.6 | Time to event modified toxicity probability interval

Time to event modified toxicity probability interval (TITE-mTPI2) model-assisted approach proposed by Lin and Yuan[8] is a time to event extension of the modified toxicity probability interval 2 design also known as the "Keyboard" design, whereby dose escalation is guided by the target interval $(\tau_1, \tau_2) = (\tau - \epsilon_1, \tau + \epsilon_2)$ where τ is the target probability of DLT. The interval [0, 1] is divided into equally sized "keys" of size $\epsilon_1 + \epsilon_2$ (apart from the end keys) and the key that has the largest probability guides whether the dose is escalated or de-escalated. These probabilities are calculated based on "effective" binomial data, including the patients whose full observation period has not yet completed. Again, the decision table can be calculated before the beginning of the trial, based on the target interval and the Beta(1, 1) prior assigned to the probability of DLT at each dose. The target interval used by the authors is $(\tau - 0.05, \tau + 0.05)$. The final dose recommendation is based on an isotonic regression once the follow-up for all enrolled patients has completed.

2.7 | Rolling modified toxicity probability interval

The rolling modified toxicity probability interval (R-mTPI2) model-assisted approach introduced by Guo et al[7] is an extension the rolling 6 design and of the modified toxicity probability interval design, whereby dose escalation is guided by the target interval $(\tau_1, \tau_2) = (\tau - \epsilon_1, \tau + \epsilon_2)$ where τ is the target probability of DLT. Again, the interval [0, 1] is divided into equally sized "keys" of size $\epsilon_1 + \epsilon_2$ (apart from the end keys). In the traditional mTPI2 design, the key that has the largest probability guides whether the dose is escalated or de-escalated, whereas in this rolling version, the decision to escalate is based on a series of rules. These rules first consider the escalation decision based solely on the number of observed DLTs within completely observed patients, then consider decisions based on the best/worst case scenarios that none/all incomplete observations are DLTs, and finally a consideration of how many patients have been consecutively assigned to the current dose without interruption. Like the other two model-assisted methods, a decision table can be calculated before the start of the trial, based on the target interval and the Beta(1, 1) prior assigned to the probability of DLT at each dose. The target interval used by the authors is $(\tau - 0.05, \tau + 0.05)$. The final dose recommendation is based on an isotonic regression once the follow-up for all enrolled patients has completed.

3 | SETTING

We consider the motivating example of the phase I trial, "First-in-human study of BAY2287411 injection, a Thorium-227 labeled antibody-chelator conjugate, in patients with tumors known to express mesothelin",[15] an ongoing trial which started in June 2018. Subjects with either advanced recurrent epithelioid mesothelioma or serous ovarian cancer who have exhausted available therapeutic options are given a single dose of Thorium-227 on day 1 of each cycle lasting 6 weeks. The dose starts at 1.5 MBq and increases in steps of 1.0 or 1.5 MBq.

In this example trial, the follow-up for observation of DLTs is only the first cycle of treatment, up to day 43, with a target DLT rate of 0.3. We base the setting for our simulation study on the motivating example, exploring the impact of considering later onset toxicities in the design. Six doses of therapy are investigated, of quantity 1.5, 2.5, 3.5, 4.5, 6.0, 7.0 MBq. The study enrolls a maximum of 30 patients in cohorts of size 3. Patients are followed up for a total of 3 treatment cycles, each of length 6 weeks. If a DLT response is observed, that patient goes off study. The first cohort of patients are always enrolled at the lowest dose.

In order to review the performances, we consider two settings, the first with minimal stopping rules, and the second with a set of stopping rules that are used as standard in such studies in practice. The first setting eliminates many confounders that may mask some aspects of the differences in performance of the methods, and the second setting is to

mimic more closely a real study. The following enforcement and stopping rules are considered, although individual numbers can be adapted according to the study itself, and any additional rules may of course be used in practice, for example personalized to patient requests. For any given dose, p_s is the $P(\text{DLT})$ in cycles up to and including cycle s.

Enforcement rules:

1. *Hard safety*: If there is a high probability that the dose exceeds the target, that dose and all above is excluded from further experimentation (ie, A dose is excluded when $P(p_1 > \tau) > threshold$). Here we use a threshold for excessive toxicity of 95%, with a Beta(1,1) prior, which translates to the following numbers. For any given dose, in the first cycle, if there are at least 3 DLT responses out of 3 patients, or at least 4 DLT responses out of 6 patients, or at least 5 DLT responses out of 9 patients, then all dose assignments must be lower than that dose for the rest of the study (ie, A dose is excluded when $P(p_1 > 30\%) > 0.95$).
2. *K-fold skipping doses*: No more than a 2-fold-rise in dose based on the highest experimented dose so far.

Stopping rules:

1. *Sufficient information*: If a dose is recommended on which 9 patients have already been treated, the trial is stopped.
2. *Lowest dose deemed unsafe*: $P(p_1 > 30\%) > 0.80$ for dose d_1 and at least one cohort of patients has been assigned to dose d_1.
3. *Highest dose deemed very safe*: $P(p_1 \leq 30\%) > 0.80$ for dose d_J and at least one cohort of patients has been assigned to dose d_J.
4. *Precision*: Stopping when MTD is precisely estimated, $CV(MTD) < 30\%$. The coefficient of variation is calculated as an adjusted median absolute deviation divided by the median. This stopping rule is only used once at least 9 patients have at least one cycle of treatment (on any dose).
5. *Hard safety*: The lowest dose is considered unsafe according to the hard safety enforcement rule.
6. *Maximum patients*: The maximum number of patients ($N = 30$) have been recruited.

In setting 1, we only consider the enforcement rule of no k-fold dose skipping, and the stopping rules of sufficient information and maximum patients. In setting 2, all enforcement and stopping rules are applied. Note that the model-assisted methods are unable to stop for precision since they only consider discrete dose levels with no model relating dose value and response. Although stopping rules 2 and 5 both stop the trial for safety concerns, it is important to highlight that they do so in a different manner. The hard safety stopping rule only considers the lowest dose in isolation, and does not use the analysis from the design itself. The lowest dose deemed unsafe rule uses the full observed data and the method of analysis from the design itself. Stopping rules 2 and 3 are implemented for the model assisted designs using the assisting Beta-binomial model, and for TITE-CRM and TITE-CRM2 by using an additional model which restricts the total follow up time to the length of one cycle.

4 | PRIOR CALIBRATION

The value of hyper-parameters of the prior distributions can have a substantial effect on the dose escalation. In a clinical setting, these can reflect belief of the toxicity of the doses, but they also have a key role in the safe and controlled escalation procedure. In order for a fair comparison between these methods, we use a calibration procedure, in line with that used by Mozgunov et al,[16] where further details to supplement the outline we present here can be found.

This calibration procedure is conducted as follows. For any given design, a grid search is performed over values of the hyper-parameters in order to find the combination of hyper-parameter values that gives the best overall performance. This performance is measured as the geometric mean of proportion of correct recommendations of MTD in 1000 simulations across a small set of clinically plausible settings. The geometric mean is used to penalize a very poor performance in any given scenario. This calibration procedure gives each design the same opportunity to achieve a good performance.

The priors are calibrated separately for setting 1 and setting 2, as setting 2 includes safety stopping rules so we must consider performance in scenarios where all doses are unsafe and where all doses are too safe. In setting 1, we cannot calibrate using such scenarios as there is no "correct" outcome.

The following scenarios in Table 1 are considered for prior calibration. In setting 1, P.S.1-P.S.4 are used, and in setting 2 we introduce P.S.5 and P.S.6, to reflect the addition of the stopping rules. In scenario P.S.5, a "correct" outcome is

TABLE 1 The P(DLT) in cycle 1 for the six doses in the six scenarios used in the prior calibration procedure, with the MTD highlighted in boldface

	p_1					
	1.5 MBq	**2.5 MBq**	**3.5 MBq**	**4.5 MBq**	**6.0 MBq**	**7.0 MBq**
P.S.1	**0.300**	0.400	0.450	0.500	0.550	0.600
P.S.2	0.050	0.070	0.100	0.150	0.200	**0.300**
P.S.3	0.100	0.200	**0.300**	0.400	0.500	0.600
P.S.4	0.150	0.200	0.250	**0.300**	0.350	0.400
P.S.5	0.400	0.450	0.500	0.550	0.600	0.650
P.S.6	0.070	0.090	0.110	0.130	0.150	0.170

stopping according to stopping rules 2 or 5, not recommending any dose. In scenario P.S.6, a "correct" outcome is stopping according to stopping rule 3.

The calibrated hyper-parameter values for the prior distributions are given in Table 2. Details of the grid over which the search was conducted are available in the supplementary materials. In many methods these values are the same in setting 1 and setting 2 (TITE-CRM, TITE-CRM2, POMM, TITE-mTPI2, R-mTPI2) and for the other methods they are relatively similar. There are some differences between these values and those in the original proposals, discussed further in Section 5.2. We also include the value of the prior effective sample size (ESS)[17] for each of the priors in Table 2. The prior ESS has a large impact on the dose escalation[18,19] and hence is important to keep low enough so as to not dominate the actual trial data. For each of the priors, this is approximately 1 to 2 patients per dose.

5 | SIMULATION STUDIES

In this section, we detail the comparative simulation studies undertaken. We first describe the data generation used in the simulation studies in Section 5.1, then present and analyze the results of those studies in Section 5.2.

5.1 | Data generation

Since the methods require different levels of information of patient response, the mechanism of generation of patient responses is not immediately obvious. Here we describe the process of generating data in a generic way. For notational simplicity, we describe the generation for a single dose and so have omitted any index referring to dose. Note that here we only describe the data-generation process so that the required data for each method are generated consistently, and not any assumptions of the analysis of each method.

Data are generated for each patient in the following way. Each patient i has a latent toxicity variable z_i drawn from a Uniform(0,1) distribution. This variable determines the outcome of patient i on all cycles and all doses.

In the data generation, it is assumed that there is a constant decrease in P(DLT) across cycles, this value is taken to be 1/3. This is reflective of cumulative toxicity. Extending the notation of defining p_s as the total true P(DLT) in cycles up to and including cycle s, we define p as the total true P(DLT) in the entire follow up period. We obtain the following for a follow up period of 3 cycles:

$$p = p_3 = p_1 + (1 - p_1)\frac{p_1}{3} + (1 - p_1)\left(1 - \frac{p_1}{3}\right)\frac{p_1}{9}.$$

For example, given a toxicity target of $p_1 = 0.3$ in the first cycle, we obtain a cumulative toxicity target across three cycles of $p_3 = 0.391$.

To generate the binary variable $Y_{i,s}$, which equals 1 if patient i observes a DLT response in cycle s and 0 otherwise, the simple indicator is used $Y_{i,s} = \mathbb{I}[p_{s-1} < 1 - z_i < p_s]$ where $p_0 = 0$ and $Y_{i,1}$ is always defined, with $Y_{i,s}$ only defined for $s > 1$ if $Y_{i,s-1} = 0$.

TABLE 2 The values of hyper-parameters resulting from the prior calibration procedure

	Setting 1	Setting 2
TITE-CRM	$\sigma^2 = 1$	$\sigma^2 = 1$
	$d = (0.05, 0.10, 0.15, 0.20, 0.25, 0.30)$	$d = (0.05, 0.10, 0.15, 0.20, 0.25, 0.30)$
	ESS ≈ 2	ESS ≈ 2
TITE-CRM2	$\mu_{a_0} = -1$	$\mu_{a_0} = -1$
	$\sigma_{a_0}^{-2} = 0.3$	$\sigma_{a_0}^{-2} = 0.3$
	$\mu_{a_1} = \log(0.2)$	$\mu_{a_1} = \log(0.2)$
	$\sigma_{a_1}^{-2} = 0.3$	$\sigma_{a_1}^{-2} = 0.3$
	ESS ≈ 1	ESS ≈ 1
ICSDP	$\pi *_1 = 0.2$	$\pi *_1 = 0.2$
	$\pi *_J = 0.4$	$\pi *_J = 0.3$
	$n_0 = 6$	$n_0 = 4$
	ESS ≈ 2	ESS ≈ 1
POMM	$p_1 * = (0.15, 0.20, 0.25, 0.3, 0.35, 0.40)$	$p_1 * = (0.15, 0.20, 0.25, 0.3, 0.35, 0.40)$
	$n_0 = 2$	$n_0 = 2$
	$p_1^{G2}/p_1 = (0.20, 0.30, 0.40, 0.50, 0.60)$	$p_1^{G2}/p_1 = (0.20, 0.30, 0.40, 0.50, 0.60)$
	ESS ≈ 2	ESS ≈ 2
nTTP	$\mu_{\beta_0} = 0.1$	$\mu_{\beta_0} = 0.05$
	$\sigma_{\beta_0}^2 = 100$	$\sigma_{\beta_0}^2 = 10$
	$\mu_{\beta_1} = 0.5$	$\mu_{\beta_1} = 0.1$
	$\sigma_{\beta_1}^2 = 100$	$\sigma_{\beta_1}^2 = 10$
	$\mu_{\beta_2} = 0$	$\mu_{\beta_2} = 0$
	$\sigma_{\beta_2}^2 = 10$	$\sigma_{\beta_2}^2 = 10$
	ESS ≈ 1	ESS ≈ 1
TITE-BOIN	$\tau_1 = 0.3128$ (equivalent to $\lambda_e = 0.3512$)	$\tau_1 = 0.3128$ (equivalent to $\lambda_e = 0.3512$)
	$\tau_2 = 0.5083$ (equivalent to $\lambda_d = 0.4492$)	$\tau_2 = 0.5083$ (equivalent to $\lambda_d = 0.4492$)
	$\alpha = 0.1$	$\alpha = 1$
	$\beta = 0.9$	$\beta = 1$
	ESS ≈ 1	ESS ≈ 2
TITE-mTPI2	$\tau_1 = 0.3519$	$\tau_1 = 0.3519$
	$\tau_2 = 0.5474$	$\tau_2 = 0.5474$
	ESS ≈ 2	ESS ≈ 2
R-mTPI2	$\tau_1 = 0.3519$	$\tau_1 = 0.3519$
	$\tau_2 = 0.5474$	$\tau_2 = 0.5474$
	ESS ≈ 2	ESS ≈ 2

Note: This corresponds to a target toxicity of 0.391 over 3 cycles. ESS is the average effective sample size per dose level.

Both the POMM and the nTTP approaches need more detailed patient responses than the binary variable $Y_{i,s}$, and so we must also generate the grades and types of toxicities. The nTTP method specifies three type of toxicity (renal, hematological, neurological) of grades 0 to 4. Patients may observe toxicities of different types, and the maximum grade is used in the POMM approach. As is standard in such studies, we classify a toxicity of grade 3 or above as a DLT.

We first illustrate how the maximum observed grades are calculated for the first cycle. First, define the probability of a grade g toxicity being the maximum observed grade in cycle 1 as p_1^{Gg}. Then let

$$
\begin{aligned}
p_1^{G4} &= p_1^{G3} = p_1/2, \\
p_1^{G2} &= (1-p_1)\mathbb{I}\left[p_1 > \frac{1}{2}\right] + p_1\mathbb{I}\left[p_1 \leq \frac{1}{2}\right], \\
p_1^{G1} &= (1-2p_1)\mathbb{I}\left[\frac{2}{5} < p_1 < \frac{1}{2}\right] + \frac{p_1}{2}\mathbb{I}\left[p_1 \leq \frac{2}{5}\right], \\
p_1^{G0} &= \left(1 - \frac{5p_1}{2}\right)\mathbb{I}\left[p_1 \leq \frac{2}{5}\right].
\end{aligned}
$$

In the same way as $Y_{i,1}$ is calculated, the observed grade $Y_{i,1}^{Gg}$ is determined by the latent variable z_i in the following way:

$$
\begin{aligned}
Y_{i,1}^{Gg} &= \mathbb{I}\left[\sum_{k=g+1}^{4} p_1^{Gk} < 1 - z_i < \sum_{k=g}^{4} p_1^{Gg}\right] \quad \text{for } g < 4, \text{ and} \\
Y_{i,1}^{G4} &= \mathbb{I}\left[1 - z_i < p_1^{G4}\right].
\end{aligned}
$$

To illustrate this, we provide an two examples, one dose where the probability of DLT in cycle 1 is exactly on target at 0.3, and a second dose where the probability of DLT in cycle 1 is 0.6.

Example 1.

$$
\begin{aligned}
p_1^{G4} &= p_1^{G3} = 0.15, \\
p_1^{G2} &= 0.3, \\
p_1^{G1} &= 0.15, \\
p_1^{G0} &= 0.25.
\end{aligned}
$$

Example 2.

$$
\begin{aligned}
p_1^{G4} &= p_1^{G3} = 0.3, \\
p_1^{G2} &= 0.4, \\
p_1^{G1} &= 0, \\
p_1^{G0} &= 0.
\end{aligned}
$$

Next, we describe how we calculate the combination of grade and type of observed toxicity. For the five grades and three types, there are therefore 125 combinations of type and grade observations. These are partitioned into sets defined by the maximum grade, with the probability that the observation is in a given set being the previously defined p_1^{Gg}. Within each set, each combination has equal probability. *nTTP* values are then calculated according to the weights given by Yin et al.[12]

For subsequent cycles, the same approach is taken, with all p_s^{Gg} values scaled accordingly, so that $p_2^{Gg} = (1-p_1)p_1^{Gg}$ and $p_3^{Gg} = (1-p_2)p_1^{Gg}$. This means that for patients with $1 - z_i < p_s$, no responses are defined for cycles after cycle s, since the patient has left the study. In this framework of generating data, the probability of each outcome is defined by the value of p_1.

5.2 | Results

Five thousand simulations are conducted for each approach across a wide range of scenarios. A full study of 17 scenarios is undertaken in order to explore the behavior of the different methods in both setting 1 and 2 with a target, τ, of the P(DLT) across the three cycles of 0.391, which although may seem high, equates to P(DLT) in cycle 1 of 0.3 in our data generation and hence is in line with a standard setting. While we focus on the results of four of these scenarios here, the full specification and results of all 17 scenarios is given in the online supplementary materials, due to space limitations. We also look at an additional set of similar scenarios in setting 2, with a target, τ, of the P(DLT) across the three cycles of 0.25. This is to investigate any differences that may occur for a lower target.

The six scenarios we look at in depth are specified in Figure 1. We consider scenarios A-D in settings 1 and 2 with a target of 0.391 across the three cycles, and scenarios C-F in setting 2 with target 0.25 across the three cycles. The MTD in scenarios A, B, D, E, and F is highlighted by the dotted line. In scenario A, the lowest investigated dose is the MTD, with p_1 linearly increasing with dose level. In scenario B, the fourth dose level is the MTD, with a nonlinear increase in p_1 by dose level. In both of these scenarios, the MTD has the exact target p_1. In scenario C, all doses are unsafe. In scenario D, the lowest three doses have $p_1 = 0.05$ and the highest three doses have $p_1 = 0.8$, hence the third dose level is the MTD. Although p_1 is clearly well below the target, the fourth dose is very unsafe and so the third dose level is by definition the maximum dose that is on target or below. In scenario E, we have a similar pattern to scenario A, but for a lower target P(DLT). Again for scenario F, there is a similar pattern to scenario B but for the lower target rate.

We use a number of metrics to compare the performance of the designs. The first is the proportion of correct selections (PCS), defined as the proportion of simulations that make the correct choice, be that recommending the true MTD or stopping for safety when all doses are unsafe. Note that in the supplementary materials, where we also consider scenarios in which all doses have P(DLT) below the target τ, we define a correct outcome in this case as stopping the trial for stopping rule 3, highest dose deemed very safe. We compare this PCS value to an empirical optimal benchmark,[20] whereby each individual patient's latent toxicity variable z_i determines that patient's response in all cycles on all dose levels. This is evaluated for all patients, as we know the response of any patient at any possible dose level and the dose with the mean response closest to the target (either nTTP or P(DLT) across the three cycles) is chosen as the recommended dose. The benchmark level is then the number of simulations that correctly identify the MTD with this full information on all patients and all doses. It is important to note that the way the simulations are conducted means that it is the same sequence of patients with the same latent toxicity variables in the simulations for all methods and the benchmark comparator. The only difference is that the benchmark always uses the maximum number of patients, whereas the other methods have the options to stop before this maximum is reached. The purpose of this benchmark is to give an indication of the difficulty of the scenario, with more difficult scenarios exhibiting a lower percentage of PCS. This metric is only concerned with dose selection and therefore we must also consider other metrics, especially those related to safety of patients within the trial.

We use two measures of the size of the trial: the total number of patients and the total length of the trial in weeks until all recruited patients have finished their follow up time or experienced DLTs. It is desirable to have a shorter trial with fewer patients. We consider the allocations of patients, focusing on the number of patients assigned to the MTD and to unsafe doses. The reason for stopping the trial is also of interest, especially in setting 2 where there are many stopping rules implemented.

In scenario A, where the lowest dose is the MTD, the best performing design in setting 1 in terms of PCS is the TITE-CRM2 (87%), closely followed by the nTTP (82%) and TITE-CRM (80%), these are also the designs with the highest allocation to the true MTD (a mean of 9, 8, and 8, respectively). These designs even outperform the benchmark, a phenomenon possible due to the sufficient information stopping rule. The worst performing design in this case is the POMM (41%), where a large number of patients are assigned to unsafe doses, on average 14. However in scenario B, where the fourth dose level is the true MTD, the POMM is the best performing design in setting 1 with a PCS of 62%. This could also potentially be driven by the prior pseudo data that in this case closely matches the true scenario for the lowest four out of the six investigated doses. The interval censored survival approach (ICSDP) is the next best performing with a PCS of 47%, with the other designs all below 30%. In this scenario we also see that both variants of the mTPI2 design have a much longer trial duration, nearly twice as long as the model-based designs.

In scenario C, where all doses are unsafe, setting 1 does not allow for stopping for safety (this is left for setting 2), therefore there is no measure of PCS. However we can still note that the POMM has the largest mean sample size of 20 patients and mean duration of 52 weeks. We can also see in Figure 2, that due to the lack of stopping rules in setting 1, the shape of the graphs for scenario A and C are very similar. In scenario D, where the third dose level is the MTD, the model-assisted methods show superior performance in terms of PCS, ranging from 83% to 93%. There is no noticeable

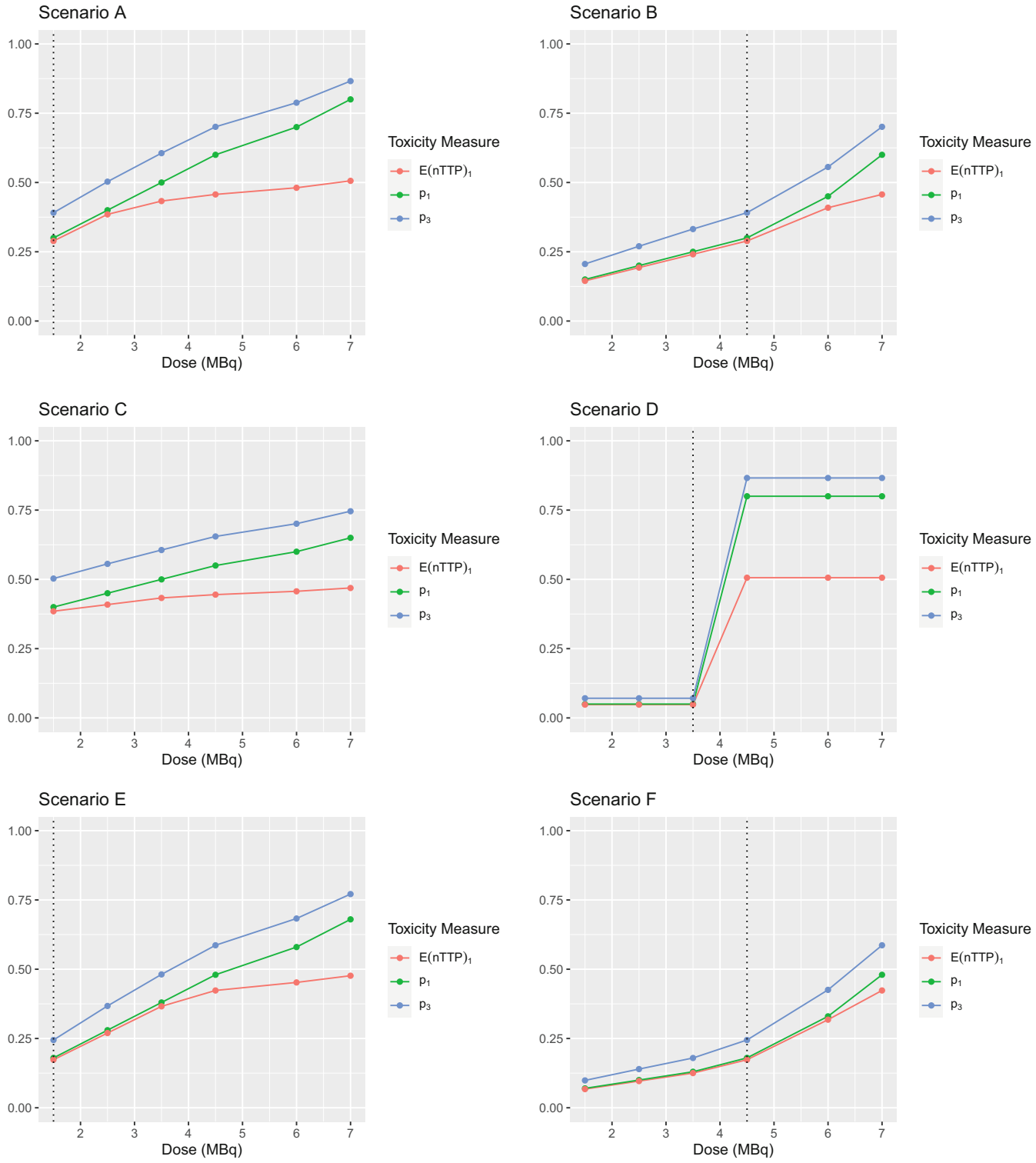


FIGURE 1 Specification of the four scenarios, with MTD highlighted in with a dotted line for scenarios A, B, D, E, and F. $E(\text{nTTP})_1$ is the expected nTTP value for cycle 1

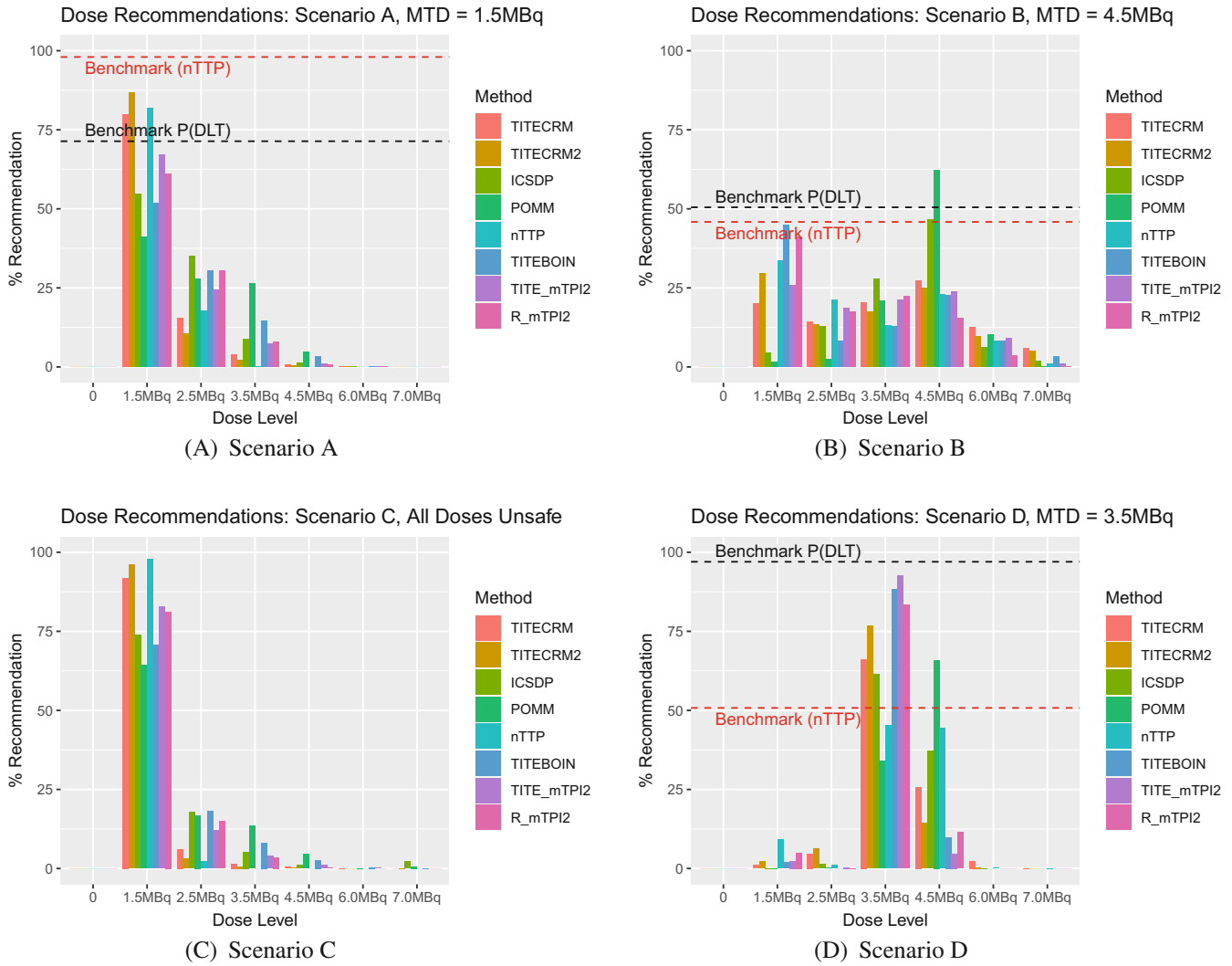
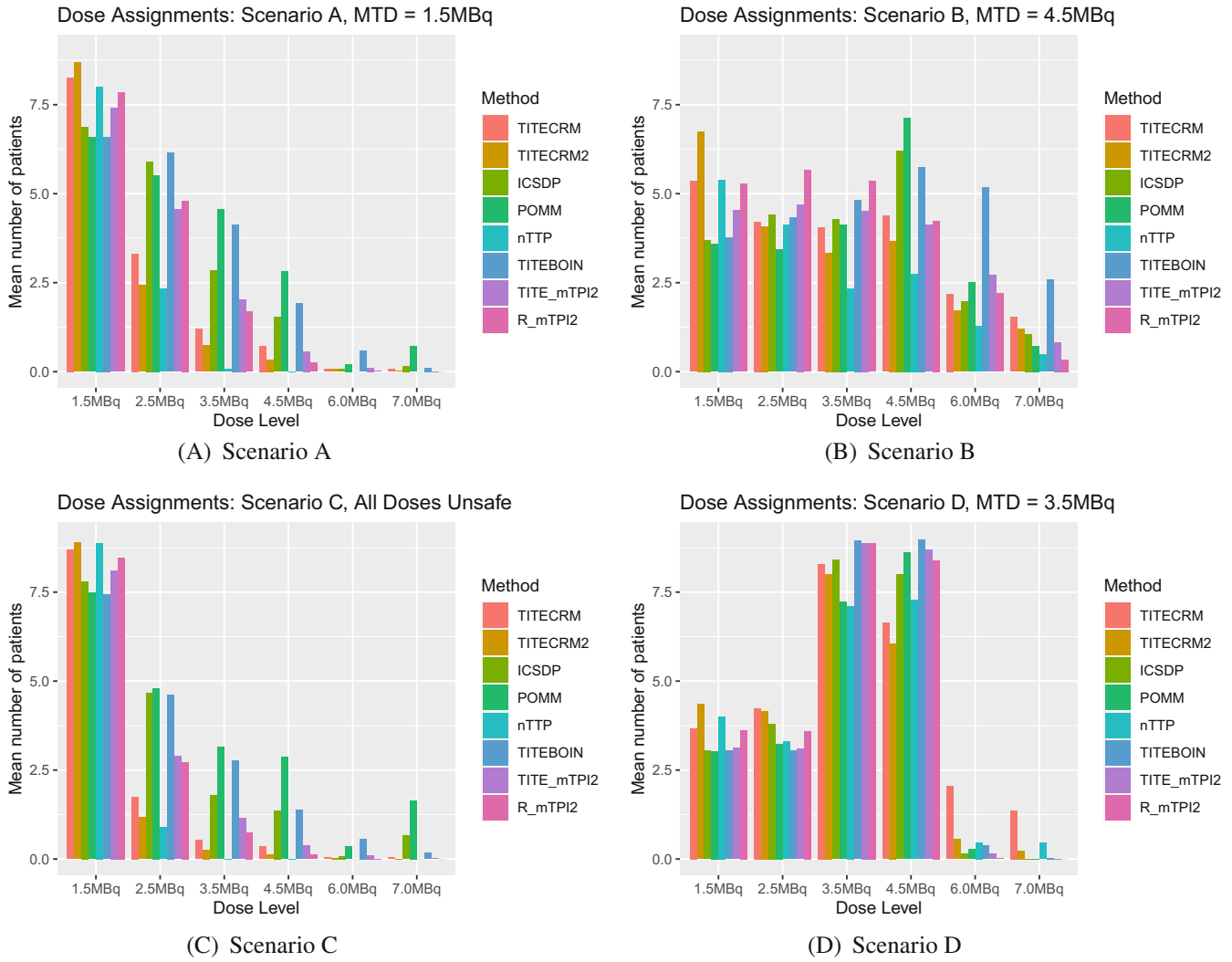


FIGURE 2 Setting 1 dose recommendations expressed as a percentage of simulations which recommend the given dose level. 0 indicates no dose is recommended, which is not applicable in setting 1

difference in sample size between these designs and the model-based ones, but there is still a large increase in the mean trial duration for the mTPI2 based designs. Figure 3 shows that for most methods, there is a similar level of assignment to the third and fourth doses, despite the vast differences in underlying toxicity between the two doses.

In setting 1, minimal stopping rules are implemented in order to investigate the behavior of the designs by themselves. In setting 2 however, the rules are more reflective of a true dose-finding trial, with stopping rules for safety and precision. In scenario A, this is especially noticeable in that it introduces the extra possibility of wrongly stopping the trial because the lowest dose is deemed unsafe. Comparing Figure 2A to Figure 4A, it is clear that for all methods, a large proportion of simulations that in setting 1 correctly recommended the lowest dose as the MTD, now stop early for safety. This is especially prevalent for the TITE-CRM2 method, the best performing in setting 1, where 50% of simulations are stopped for this reason. The best performing method in scenario A in setting 2 is the nTTP, although a PCS of 57% in a scenario where the lowest dose is the MTD is by no means an outstanding performance.

Scenario B shows less of a contrast in setting 2 to setting 1, with POMM and ICSDP giving the best, and similar performances. In scenario C, the safety stopping rules are implemented more effectively in some methods than others. Both the TITE-CRM2 and nTTP stop for safety in around 72% to 73% of simulations, whereas the ICSDP only stops for safety in 29% of simulations, with 50% of simulations recommending the lowest dose as the MTD. There is an average of 14 patients, nearly 5 cohorts, an unacceptable level when all doses are unsafe. This is most likely driven by the prior for this method, as there are 5 pseudo-patients on the lowest dose, providing evidence that although chosen by the calibration



(A) Scenario A (B) Scenario B

(C) Scenario C (D) Scenario D

FIGURE 3 Setting 1 dose assignments expressed as an average number of patients over simulations which are assigned the given dose level

procedure, such a prior may be too strong to use practically. The POMM also has an average of 14 patients in this scenario, although stops in 50% of simulations. The dose assignment for this method is more unsafe than the ICSDP, indicated in Figure 5C, by the high levels of assignment to higher dose levels. The TITE-BOIN design also sees high levels of assignment to higher doses. In general we see the model-assisted designs allocating higher numbers of patients to unsafe doses due to the asymmetric target interval resulting from the prior calibration. Results for the model assisted designs with the originally proposed prior hyper-parameters are available in the supplementary materials, illustrating that the lower number of patients assigned to unsafe doses is accompanied by a much lower PCS in scenarios where the MTD is higher in the dose range (Tables 3-6).

In scenario D, the PCS is improved in setting 2 over setting 1, with four methods (ICSDP, TITE-BOIN, TITE-mTPI2, and R-mTPI2) achieving the benchmark level and nTTP even exceeds this. This is in part due to the hard safety rule that eliminates unsafe doses, and in part to the precision stopping rule. The dose assignment shows this is the case, comparing Figure 5D to Figure 3D clearly shows the reduction in assignment to the fourth dose.

When considering a lower target DLT rate in setting 2, Figure 6 shows the dose recommendations, Figure 7 shows the dose assignments and Table 7 gives the mean duration and size of the trials in scenarios C-F. In general we see very similar patterns to the corresponding scenarios with the higher target DLT rate, with a few small differences. In scenario C, where all doses are unsafe, fewer patients are assigned, as expected when the difference between the target and the toxicity of the lowest dose is larger. In terms of dose recommendations, performance is generally better, despite prior

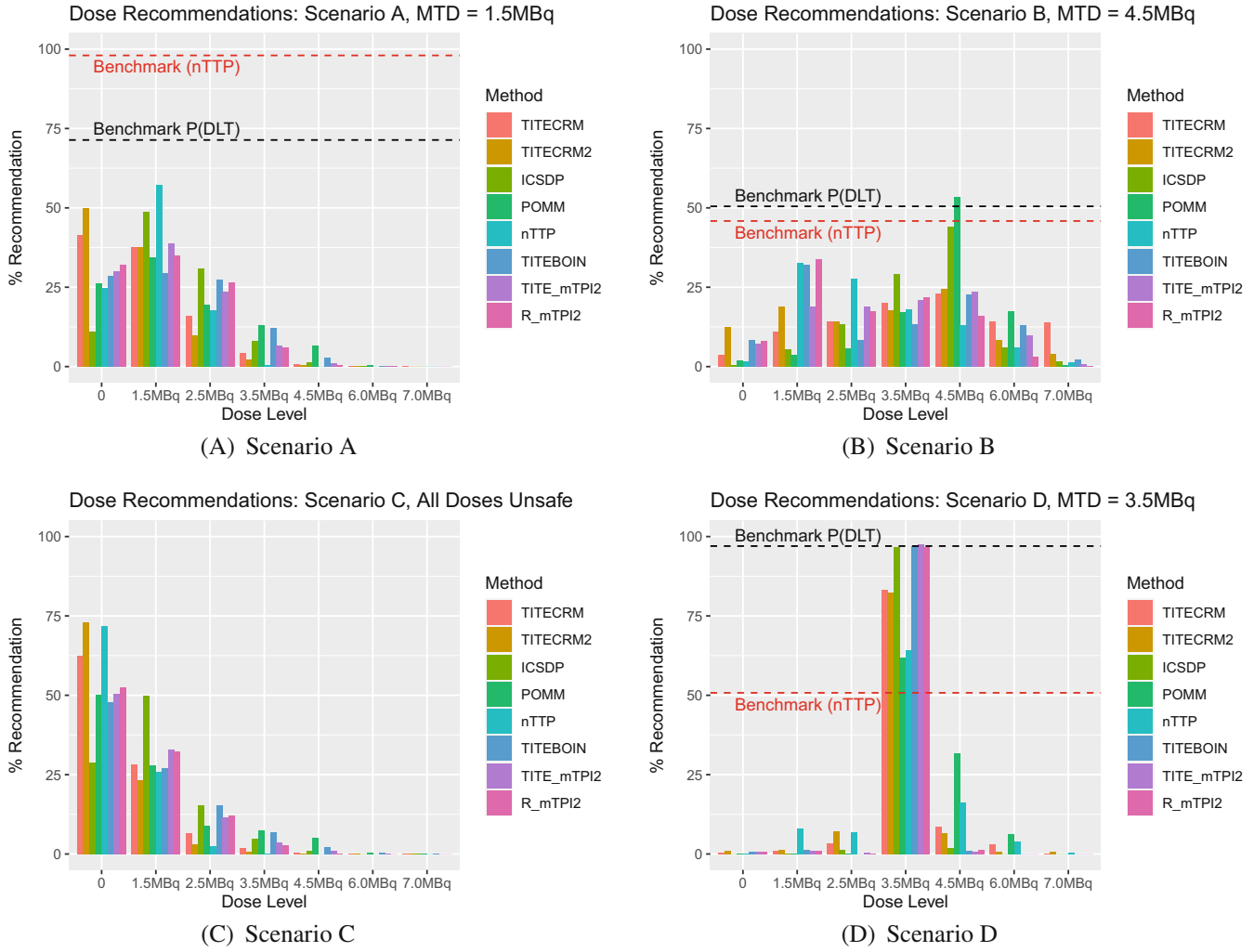
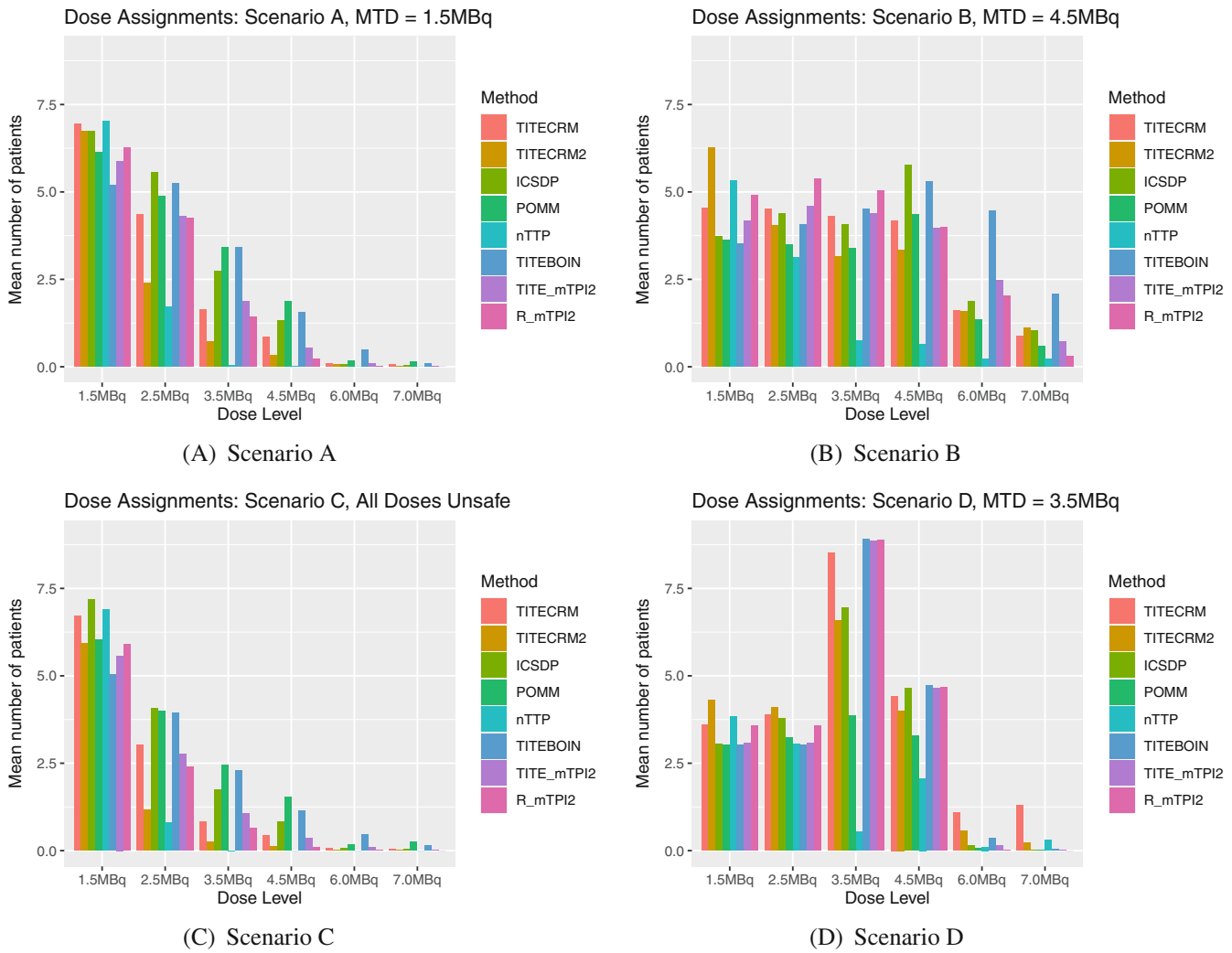


FIGURE 4 Setting 2 dose recommendations for a target of $\tau = 0.391$ expressed as a percentage of simulations which recommend the given dose level. 0 indicates no dose is recommended

specifications having been calibrated using scenarios with the higher target rate. The exception being the nTTP method which performs considerably poorer in scenario D, due to the nonlinearity of the true dose-response relationship.

Observing the overall performance across the entire range of scenarios, the front-runner is the ICSDP, despite its poor performance when all doses are unsafe in setting 2, especially for the lower target DLT rate. The next best performing method is TITE-CRM2, which gives a good yet balanced performance across scenarios in both settings. The TITE-CRM2 also has a slightly shorter average trial duration, requiring fewer patients overall and fewer patients treated on unsafe doses. However we also note that the process of data generation follows more closely the ICSDP assumptions than the TITE-CRM2, and so this comparison may be slightly different under the different assumptions.

In both settings, the mTPI2 based methods clearly have a longer than ideal trial length, as evidenced in the full study of 17 scenarios where their average durations are nearly twice those of the other methods, and hence would not be recommended for use. The TITE-BOIN also has larger patient numbers in most scenarios. It is important to note that the relationship between trial duration and the total trial size for the methods varies due to the rules implemented in the different methods. The model-assisted approaches' PCS are also less than the model-based approaches in most scenarios in setting 1, although this is improved somewhat in setting 2. Of the model-based approaches, although the nTTP provides the shortest trial duration, the PCS is much lower than the other designs when the true MTD is in the higher doses in setting 2, and so also not recommended. This is mainly due to premature stopping for precision, a consequence of the fact that this rule is designed for binary rather than continuous endpoints.



(A) Scenario A (B) Scenario B

(C) Scenario C (D) Scenario D

FIGURE 5 Setting 2 dose assignments for a target of $\tau = 0.391$ expressed as an average number of patients over simulations which are assigned the given dose level

TABLE 3 Setting 1: Measures of size of the trial across scenarios, total duration in weeks, and total number of patients

	TITE-CRM	TITE-CRM2	ICSDP	POMM	nTTP	TITE-BOIN	TITE-mTPI2	R-mTPI2
Scenario	*Mean duration in weeks (SD)*							
A	39 (11)	36 (10)	46 (10)	52 (9)	32 (5)	40 (14)	46 (23)	50 (26)
B	55 (14)	53 (15)	55 (10)	55 (8)	44 (14)	57 (13)	81 (31)	95 (34)
C	34 (9)	32 (7)	44 (10)	52 (9)	31 (4)	35 (16)	37 (22)	37 (23)
D	63 (7)	57 (8)	57 (5)	55 (6)	56 (10)	49 (3)	85 (7)	100 (9)
Mean	48	45	51	53	41	45	62	70
Scenario	*Mean number of patients (SD)*							
A	14 (6)	12 (5)	17 (5)	20 (4)	10 (2)	19 (6)	15 (5)	15 (5)
B	22 (7)	21 (8)	22 (5)	22 (4)	16 (7)	26 (4)	21 (6)	23 (7)
C	11 (4)	10 (3)	16 (5)	20 (4)	10 (2)	17 (7)	13 (5)	12 (5)
D	26 (4)	23 (4)	23 (3)	22 (2)	23 (5)	24 (1)	24 (2)	25 (2)
Mean	18	17	20	21	15	22	18	19

TABLE 4 Setting 1: Summary of stopping reasons for setting 1, expressed as a percentage of simulations where the given stopping rule was triggered

	TITE-CRM	TITE-CRM2	ICSDP	POMM	nTTP	TITE-BOIN	TITE-mTPI2	R-mTPI2
Stopping reason	*Scenario A*							
Sufficient information	99	100	100	97	100	95	100	99
Maximum patients	2	1	2	3	0	10	1	1
Stopping reason	*Scenario B*							
Sufficient information	90	88	95	95	96	80	95	75
Maximum patients	21	22	11	5	4	41	11	32
Stopping reason	*Scenario C*							
Sufficient information	100	100	100	96	100	95	100	100
Maximum patients	1	0	1	4	0	8	1	1
Stopping reason	*Scenario D*							
Sufficient information	89	94	100	99	96	99	99	99
Maximum patients	32	16	2	1	4	2	2	5

Note: The sum of these may be greater than 100, since it is possible for more than one rule to be triggered in a single trial.

TABLE 5 Setting 2, $\tau = 0.391$: Measures of size of the trial across scenarios, total duration in weeks, and total number of patients

	TITE-CRM	TITE-CRM2	ICSDP	POMM	nTTP	TITE-BOIN	TITE-mTPI2	R-mTPI2
Scenario	*Mean duration in weeks (SD)*							
A	39 (13)	31 (13)	44 (12)	44 (13)	29 (4)	33 (18)	41 (26)	43 (30)
B	52 (12)	50 (16)	53 (10)	45 (9)	32 (6)	51 (17)	78 (32)	90 (38)
C	32 (13)	25 (11)	38 (14)	39 (15)	27 (5)	27 (19)	32 (25)	29 (27)
D	57 (9)	50 (10)	49 (5)	36 (7)	29 (5)	40 (7)	76 (10)	92 (12)
Mean	45	39	46	41	29	38	57	63
Scenario	*Mean number of patients (SD)*							
A	14 (6)	10 (6)	16 (6)	17 (6)	9 (2)	16 (8)	13 (7)	12 (7)
B	20 (6)	19 (8)	21 (5)	17 (4)	10 (3)	24 (7)	20 (7)	22 (8)
C	11 (6)	8 (5)	14 (6)	14 (7)	8 (2)	13 (9)	10 (7)	9 (6)
D	23 (5)	20 (4)	19 (3)	13 (3)	10 (2)	20 (3)	20 (3)	21 (3)
Mean	17	14	17	15	9	18	16	16

6 | DISCUSSION

In this article, we conducted a simulation study to compare the leading methods for dose-finding trials incorporating later onset toxicities in a variety of scenarios. The purpose of such a comparison was to evaluate the performance of the different methods in generic settings where their individuals assumptions may not hold, in order to highlight any key differences.

The values of the hyper-parameters for the prior distributions in each method were calculated using a calibration procedure. They were calibrated over a small number of clinically plausible scenarios, but still ranging in the position of the MTD in the dosing sequence. This ensured that the different methods all had the same opportunity to achieve their potential in a large range of scenarios. Although sensible for the purpose of this comparison, the choice of values does raise some questions for future thought. A superior performance across the calibration scenarios can give a somewhat informative prior, which can adversely affect performance in some scenarios and positively in others. The similarities of

TABLE 6 Setting 2, $\tau = 0.391$: Summary of stopping reasons for setting 2, expressed as a percentage of simulations where the given stopping rule was triggered

	TITE-CRM	TITE-CRM2	ICSDP	POMM	nTTP	TITE-BOIN	TITE-mTPI2	R-mTPI2
Stopping reason	*Scenario A*							
Sufficient information	57	46	88	60	46	68	70	68
Lowest dose deemed unsafe	42	50	11	20	22	28	30	32
Highest dose deemed too safe	0	0	0	0	0	0	0	0
Precision	0	17	2	14	73	0	0	0
Hard safety	10	4	11	14	10	3	4	4
Maximum patients	2	1	1	2	0	7	1	1
Unsafe (total)	42	50	11	26	25	28	30	32
Sufficient/precision (total)	58	50	89	71	75	68	70	68
Stopping reason	*Scenario B*							
Sufficient information	65	65	83	35	32	77	89	70
Lowest dose deemed unsafe	4	11	1	1	1	6	6	8
Highest dose deemed too safe	0	1	0	0	0	2	1	1
Precision	34	21	21	72	96	0	0	0
Hard safety	1	0	1	1	1	0	0	0
Maximum patients	7	17	9	2	0	29	8	27
Unsafe (total)	4	11	1	2	1	6	6	8
Sufficient/precision (total)	94	79	95	97	98	77	89	70
Stopping reason	*Scenario C*							
Sufficient information	37	24	70	42	20	49	49	47
Lowest dose deemed unsafe	63	73	29	38	71	48	51	53
Highest dose deemed too safe	0	0	0	0	0	0	0	0
Precision	0	19	1	6	32	0	0	0
Hard safety	24	10	29	30	21	9	10	10
Maximum patients	1	0	1	2	0	6	1	0
Unsafe (total)	63	73	29	50	72	48	51	53
Sufficient/precision (total)	37	27	71	48	28	49	49	47
Stopping reason	*Scenario D*							
Sufficient information	96	64	63	5	8	99	99	99
Lowest dose deemed unsafe	0	1	0	0	0	1	1	1
Highest dose deemed too safe	0	0	0	0	0	0	0	0
Precision	0	44	60	99	100	0	0	0
Hard safety	0	0	0	0	0	0	0	0
Maximum patients	13	4	0	0	0	1	1	1
Unsafe (total)	0	1	0	0	0	1	1	1
Sufficient/precision (total)	96	98	100	100	100	99	99	99

Note: The sum of these may be greater than 100, since it is possible for more than one rule to be triggered in a single trial.

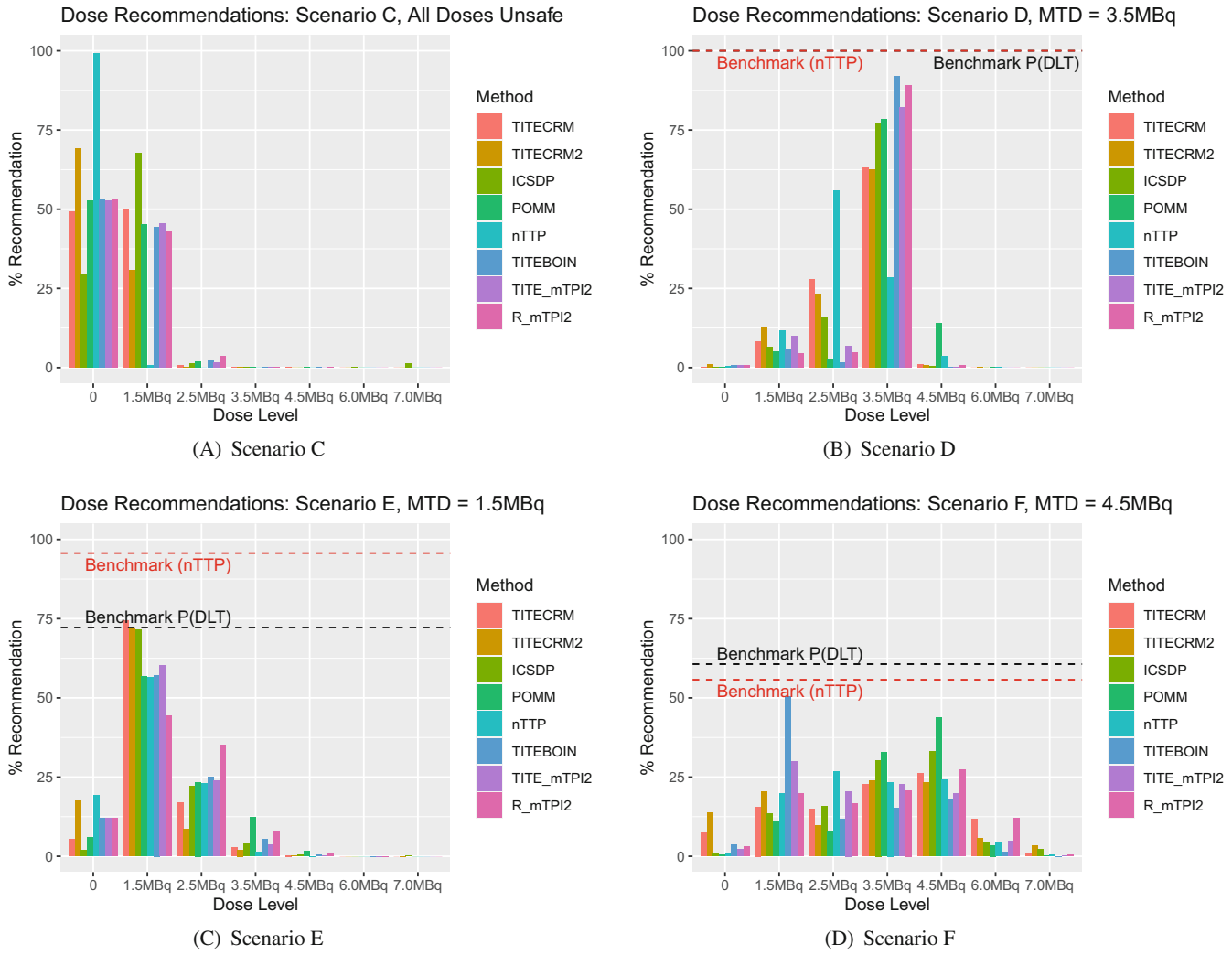


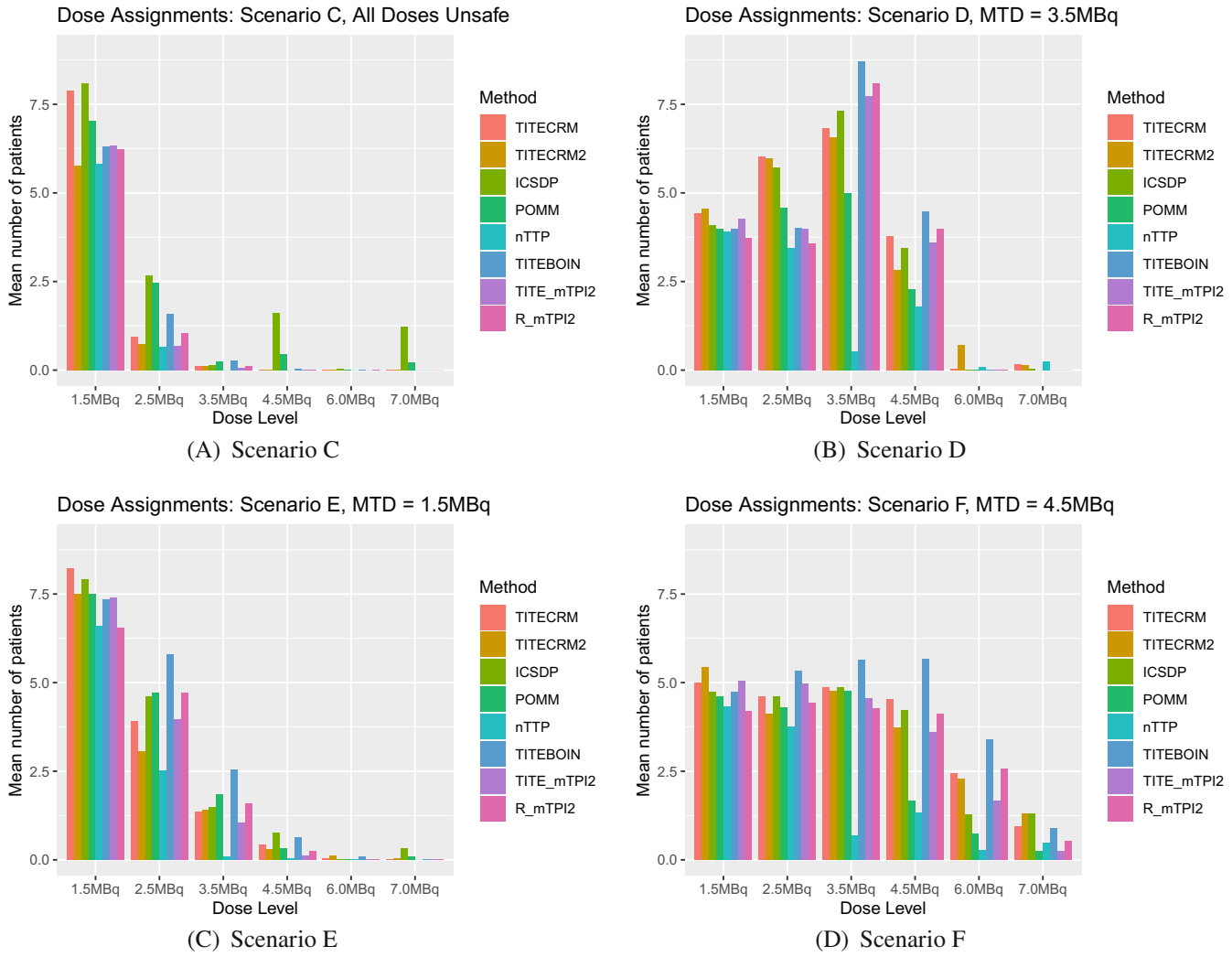
FIGURE 6 Setting 2 dose recommendations for a target of $\tau = 0.25$ expressed as a percentage of simulations which recommend the given dose level. 0 indicates no dose is recommended

the performances when the target toxicity rate is lowered demonstrates the robustness of the calibration procedure to give an overall strong performance.

Both model-based and model-assisted approaches were explored in the study, with differences in their results reflecting the difference in the methods. The model-assisted approaches offer the advantages of fewer assumptions on the dose-response relationship, which is clear to see in very strong performance in scenario D of the simulation study, where the pattern of toxicity risk does not adhere to any standard dose-response model. However, without the assistance of a model to guide the escalation, larger number of patients are on average treated on unsafe doses. This pattern is of course not unique to late onset toxicities, but is potentially accentuated by this.

It is worth noting that in order to compare the designs in a fair manner, we have had to simplify some of their features. For example, the TITE based methods are capable of including a continuous time-to-event variables as opposed to the cycle of the event. In these simulations, our simplifications have not adversely affected the comparisons, since the discretization of the time-to-event captures the behavior across the cycles, and most importantly the accrual of cohorts aligns with the discretization. If the exact timing of the event is critical for the particular drug in the trial then this of course can be incorporated in the design.

In these implementations we have assumed in our data generation that the risk of DLT decreases linearly across cycles. It may be of interest to additionally perform a sensitivity analysis on this relationship. However, careful consideration must be given to the data generation process and the definition of the target in order to correctly identify the true MTD in each scenario.



(A) Scenario C (B) Scenario D

(C) Scenario E (D) Scenario F

FIGURE 7 Setting 2 dose assignments for a target of $\tau = 0.25$ expressed as an average number of patients over simulations which are assigned the given dose level

The scenarios we investigated were purposefully difficult in order to test the methods. For example, in scenario B, there is only a 0.05 difference in toxicity risk between the 3.5 and 4.5 MBq dose. This is illustrated with the empirical benchmark only achieving 50% PCS. Results as such must be considered in the context of the scenario. In a real dose-finding trial, it is likely that the scenario may be "easier" and hence we should not be concerned by such low PCS in this case.

However, in scenario A, the low PCS in setting 2 across methods is of some concern. Setting 2 is more reflective of the implementation of a dose-finding trial, with stopping rules in place to ensure the safety of patients. In this case, these stopping rules are then overly implemented. It is perhaps worth reconsidering that in trials with late onset toxicities, the traditional safety stopping rules may be too strict.

This highlights the importance of considering the impact of the stopping rules on the different statistical methods, as they have a large effect on the true performance of any method used in practice. While it is both interesting and informative to explore the behavior of the approaches without the implementation of stopping rules, there is a limit to the usefulness of such simulations in isolation.

As well as safety stopping rules, the impact of the precision stopping rule must be considered. This is not applied to the model-assisted methods as the precision of the MTD cannot be estimated without a model. In the case of the nTTP, the precision rule is very often implemented too soon and hence the MTD is underestimated. It is therefore important to consider when using alternative measures of toxicity, whether the traditional approach to stopping rules are actually applicable.

TABLE 7 Setting 2, $\tau = 0.25$: Measures of size of the trial across scenarios, total duration in weeks, and total number of patients

	TITE-CRM	TITE-CRM2	ICSDP	POMM	nTTP	TITE-BOIN	TITE-mTPI2	R-mTPI2
Scenario	*Mean duration in weeks (SD)*							
C	28 (9)	24 (8)	38 (12)	31 (10)	24 (3)	16 (11)	22 (14)	21 (17)
D	54 (10)	53 (10)	53 (8)	43 (4)	29 (5)	43 (8)	73 (17)	92 (22)
E	40 (11)	37 (12)	42 (10)	41 (9)	30 (4)	33 (15)	42 (20)	47 (26)
F	57 (13)	55 (14)	54 (12)	44 (6)	33 (6)	56 (15)	79 (31)	86 (34)
Mean	45	42	47	40	29	37	54	61
Scenario	*Mean number of patients (SD)*							
C	9 (3)	7 (3)	14 (5)	10 (4)	6 (1)	8 (5)	7 (4)	7 (4)
D	21 (5)	21 (5)	21 (4)	16 (2)	10 (2)	21 (4)	20 (5)	19 (4)
E	14 (5)	12 (6)	15 (5)	15 (5)	9 (2)	16 (7)	13 (5)	13 (6)
F	22 (7)	22 (7)	21 (6)	16 (3)	11 (3)	26 (5)	20 (7)	20 (6)
Mean	17	15	18	14	9	18	15	15

A limitation of such a simulation study is that the decisions in a real trial that would made be on an individual patient basis cannot be incorporated. Any rule must be prespecified and therefore deviates from the practical realities of such trials. It is not our explicit intention to recommend whether or not including later cycles is beneficial or not in deciding the MTD, as this will be a decision individual to the trial in many ways, taking into account multiple factors that we cannot include in a simulation study. However, we hope that this article will provide some insight into how late onset toxicity can be incorporated if it is deemed appropriate.

ACKNOWLEDGEMENTS

This report is independent research supported by the National Institute for Health Research (Professor Thomas Jaki's Senior Research Fellowship, NIHR-SRF-2015-08-001) and by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care (DHSC). Thomas Jaki and Helen Barnett received funding from UK Medical Research Council (MC_UU_00002/14).

DATA AVAILABILITY STATEMENT

All data are simulated according to the specifications described. Software in the form of R code used to produce the presented results is available at https://github.com/helenyb/DF_Late_Safety.

ORCID

Helen Barnett https://orcid.org/0000-0001-7466-7033
Thomas Jaki https://orcid.org/0000-0002-1096-188X

REFERENCES

1. Storer BE. Phase I trials. In: Redmond C, Colton T, eds. *Biostatistics in Clinical Trials*. Chichester: John Wiley & Sons, Ltd.; 2001:337-342.
2. National Cancer Institute. Common terminology criteria for adverse events (CTCAE) v5.0; 2017. http://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm.
3. Postel-Vinay S, Gomez-Roca C, Molife LR, et al. Phase I trials of molecularly targeted agents: should we pay more attention to late toxicities? *J Clin Oncol*. 2011;29(13):1728-1735. doi:10.1200/JCO.2010.31.9236
4. Skolnik JM, Barrett JS, Jayaraman B, Patel D, Adamson PC. Shortening the timeline of pediatric Phase I trials: the rolling six design. *J Clin Oncol*. 2008;26(2):190-195. doi:10.1200/JCO.2007.12.7712
5. Storer BE. Design and analysis of Phase I clinical trials. *Biometr*. 1989;45(3):925-937.
6. Yuan Y, Lin R, Li D, Nie L, Warren KE. Time-to-event Bayesian optimal interval design to accelerate Phase I trials. *Clin Cancer Res*. 2018;24(20):4921-4930. doi:10.1158/1078-0432.CCR-18-0246.Time-to-event
7. Guo W, Ji Y, Li D. R-TPI: rolling toxicity probability interval design to shorten the duration and maintain safety of Phase I trials. *J Biopharm Stat*. 2019;29(3):411-424. doi:10.1080/10543406.2019.1577683

8. Lin R, Yuan Y. Time-to-event model-assisted designs for dose-finding trials with delayed toxicity. *Biostatistics*. 2020;21(4):807-824. doi:10.1093/biostatistics/kxz007
9. Cheung YK, Chappell R. Sequential designs for Phase I clinical trials with late-onset toxicities. *Biometrics*. 2000;56(4):1177-1182. doi:10.1111/j.0006-341x.2000.01177.x
10. Sinclair K, Whitehead A. A Bayesian approach to dose-finding studies for cancer therapies: incorporating later cycles of therapy. *Stat Med*. 2014;33(15):2665-2680. doi:10.1002/sim.6132
11. Doussau A, Thiébaut R, Paoletti X. Dose-finding design using mixed-effect proportional odds model for longitudinal graded toxicity data in Phase I oncology clinical trials. *Stat Med*. 2013;32(30):5430-5447. doi:10.1002/sim.5960
12. Yin J, Qin R, Ezzalfani M, Sargent DJ, Mandrekar SJ. A Bayesian dose-finding design incorporating toxicity data from multiple treatment cycles. *Stat Med*. 2017;36(1):67-80. doi:10.1002/sim.7134
13. Neuenschwander B, Branson M, Gsponer T. Critical aspects of the Bayesian approach to Phase I cancer trials. *Stat Med*. 2008;27:2420-2439. doi:10.1002/sim
14. Whitehead J, Williamson D. Bayesian decision procedures based on logistic regression models for dose-finding studies. *J Biopharm Stat*. 1998;8(3):445-467. doi:10.1080/10543409808835252
15. U. S. National Library of Medicine. An open-label, first-in-human, multi-center study to evaluate the safety, tolerability, pharmacokinetics and anti-tumor activity of a thorium-227 labeled antibody-chelator conjugate, BAY2287411 injection. Patients With Solid Tumors Known to Express Mes; 2018. https://clinicaltrials.gov/ct2/show/NCT03507452.
16. Mozgunov P, Knight R, Barnett H, Jaki T. Using an interaction parameter in model-based Phase I trials for combination treatments? A simulation study. *Int J Environ Res Publ Health*. 2021;18(1):1-19. doi:10.3390/ijerph18010345
17. Morita S, Thall PF, Müller P. Determining the effective sample size of a parametric prior. *Biometrics*. 2008;64(2):595-602. doi:10.1111/j.1541-0420.2007.00888.x
18. Neuenschwander B, Weber S, Schmidli H, O'Hagan A. Predictively consistent prior effective sample sizes. *Biometrics*. 2020;76(2):578-587. doi:10.1111/biom.13252
19. Wiesenfarth M, Calderazzo S. Quantification of prior impact in terms of effective current sample size. *Biometrics*. 2020;76(1):326-336. doi:10.1111/biom.13124
20. O'Quigley J, Paoletti X, Maccario J. Non-parametric optimal design in dose finding studies. *Biostatistics*. 2002;3(1):51-56. doi:10.1093/biostatistics/3.1.51

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Barnett H, Boix O, Kontos D, Jaki T. Dose finding studies for therapies with late-onset toxicities: A comparison study of designs. *Statistics in Medicine*. 2022;41(30):5767-5788. doi: 10.1002/sim.9593