学位論文

Optimal Adaptive Allocation using Deep Reinforcement Learning in Dose-Finding Studies

(用量設定試験における
被験者の割り付けの
深層強化学習による最適化)

Kentaro Matsuura
Tokyo University of Science

March, 2023

Abstract

The dose of a drug is determined based on the results of clinical trials. A dose-finding study is a clinical trial in which multiple dose candidates are administered to subjects to find the appropriate dose based on the clinical outcomes of the subjects. This paper proposes a method that efficiently allocates subjects to doses to be evaluated using deep reinforcement learning for the dose-finding studies of anticancer and non-anticancer drugs, and shows that these methods are useful in practice.

Chapter 1 provides an overview of dose-finding studies for anticancer and non-anticancer drugs, and describes the statistical issue that existing methods do not adequately optimize the performance metrics.

In Chapter 2, we propose an adaptive allocation method that optimizes a performance metric using deep reinforcement learning for dose-finding studies of non-anticancer drugs. Through numerical experiments, we compare the performance of the proposed method with existing methods and show that the proposed method improves the performance metric to be optimized. In particular, the result shows that when the mean absolute error is used as the performance metric to be optimized, the performance of other metrics is also likely to improve, and that the performance is likely to be maintained even in scenarios that deviate from the environment used for training.

In Chapter 3, we propose a dose escalation method that maximizes the percentage of correct selection of a target dose using deep reinforcement learning for dose-finding studies of anticancer drugs. Through numerical experiments, we compare the proposed method with existing methods and show that the proposed method achieves a higher percentage of correct selection than the existing methods although the it causes more toxicity. In particular, the proposed method performs better under scenarios in which the target dose is positioned at higher doses, as assumed in actual clinical trials.

Chapter 4 summarizes the limitations and challenges in applying the proposed method to actual clinical trials.

In Chapter 5, we briefly summarize the results of our study and then present the conclusions and contributions of the methods proposed in Chapters 2 and 3.

Contents

Chapter 1

Introduction

1.1 General Background of Dose-Finding Study

The probability of success from Phase I trials to approval is generally 10–20% (Smietana et al., 2016; Wong et al., 2019). One reason for this low success probability is inappropriate dose selection. Selecting a dose that is too safe will lead to a lack of efficacy and failure of late phase clinical trials. Selecting a dose that is too high will cause unnecessary adverse events. Estimating the dose-outcome relationship and selecting the appropriate dose is one of the most important steps in drug development.

A trial in which multiple doses are administered to subjects and the appropriate dose is explored based on outcomes is called a dose-finding study. In general, dose-finding studies correspond to Phase II trials for non-anticancer drugs and Phase I trials for anticancer drugs. The outcome of a Phase II trial for a non-anticancer drug is efficacy, and efficacy is confirmed in the same subject population as in a Phase III trial. The relationship between dose and efficacy response is estimated and the target dose to be used in the Phase III trials is determined. The outcome of a Phase I trial for an anticancer drug is toxicity, and toxicity is confirmed in subjects who have failed to respond to standard therapy. The relationship between dose and toxicity is estimated, and the maximum tolerated dose (MTD), which is the upper limit of the dose to be used in the later clinical trials, is estimated.

Limitations of existing methods used in these dose-finding studies include inadequate accuracy in estimating the dose-response relationship and the target dose, or the dose-toxicity relationship and MTD. This study aims to propose a method to improve the accuracy of those estimates.

1.2 Overview of Dose-Finding Studies for Non-Anticancer Drugs

In typical dose-finding studies for non-anticancer drugs, the number of subjects ranges from 50 to 300 and the number of doses ranges from 3 to 6, including the placebo group (no-dose group). For clarity, subjects are often randomized and equally assigned to each dose group. Many methods have been proposed to estimate dose-response curves (dose-response relationship), including analysis of variance (ANOVA), Bayesian modeling average (BMA)-based method (Ohlssen and Racine, 2015; Gould, 2019), multiple comparison procedure – modeling (MCP-Mod) method (Bretz et al., 2005), etc. The ANOVA is a classical analysis method to test whether the response increases with increasing

dose. The BMA methods estimate the joint posterior distribution of the dose-response curve given multiple models and the prior distributions of their shape parameters.

The MCP-Mod method is divided into a first step, MCP (multiple comparison procedure), and a second step, Mod (modeling). The following assumptions need to be made before the start of the trial: multiple candidate models for the dose-response curve (e.g., linear, emax, and sigmoid emax), shape parameters of these models, prior probabilities for each model, maximum effect in the dose range used in the trial, the variance of observations, the clinically relevant target effect, and the significance level. As discussed later, in the proposed method, these assumptions are used to generate data within reinforcement learning. Before the start of the trial, contrast coefficients for increasing power are calculated for each model. After the end of the trial, in the MCP part, T-statistics are calculated from the dose and response data to determine whether the dose-response relationship exists or not (i.e., whether the dose-response curve is non-flat or not) by multiple testing. In the subsequent Mod part, each model that passes the test is fitted to the data to estimate the shape parameters. The final model is then selected based on the AIC (or T-statistic). Based on the final model, the lowest dose that satisfies a clinically relevant target effect is used as an estimate of the target dose. The MCP-Mod method has continued to be extended after its proposal by Bretz et al. (2005) and is commonly used in recent clinical trials, partly due to the EMA guideline (EMA, 2014), the FDA guideline (FDA, 2015), and the easy-to-use R package DoseFinding. In this thesis, we mainly deal with clinical trials that use the MCP-Mod method as a method to estimate dose-response curves. However, the proposed method can be applied in the same way when using ANOVA or BMA methods.

1.3 Overview of Dose-Finding Studies for Anticancer Drugs

In typical dose-finding studies for new anticancer drugs, the number of subjects ranges from 10 to 40 and the number of doses ranges from 3 to 8. The trial is conducted per cohort, and the number of subjects in a cohort is often 3. Since this is the first time the new drug is administered to humans, the cohort is started with the lowest dose to ensure safety, and the dose for the next cohort is determined based on toxicity results. Because the dose is gradually increased from the lowest dose, this trial is also called a dose-escalation study. Toxicities that should be considered dose-limiting are called dose-limiting toxicities (DLTs), and the number of DLTs is defined as the number of subjects who experienced a DLT. Although we want to increase the dose of anticancer drugs as much as possible to increase efficacy, we also want to keep the DLT probability to a specified value. The maximum tolerated dose (MTD) is the maximum dose at which the DLT probability is within an acceptable range. Estimation of the MTD is the main purpose of the dose-finding study for a new anticancer drug (Green et al., 2012).

The most commonly used methods to estimate the MTD with dose escalation are rule-based designs such as the 3+3 design, model-based designs such as the continuous reassessment method (CRM) method, and model-assisted designs such as the BOIN method (Le Tourneau et al., 2009; Daimon et al., 2019). The 3+3 design determines an action ("continue the trial and escalate/stay/de-escalate the dose in the next cohort" or "discontinue the trial") based on the number of DLTs in the current cohort and simple rules. While it has the advantage of being easy to understand, it also has the disadvantages of not making effective use of cumulative data, a tendency to underestimate MTD, and early termination of the trial due to the accidental occurrence of DLTs in 2 out of 3 subjects. The CRM method fits an exponential model to all data available up to that time point to estimate a dose-toxicity curve, and determines an action according to the joint posterior distribution of the

DLT probabilities at each dose (O'Quigley et al., 1990). While it has the advantage of data being used effectively, it has the disadvantage of making it difficult to understand the criteria for dose escalation. The BOIN method uses a model, but the behavior is rule-based and easy to understand (Liu and Yuan, 2015). The disadvantage is that the accuracy may not be as good as the model-based method.

1.4 Statistical Issue

The issue to be focused on in this study was defined as follows:

Issue: In dose-finding studies for both non-anticancer and anticancer drugs, the allocation of subjects to each dose has not been determined based on the performance metrics to be optimized

Studies of methods in Phase II dose-finding trials for non-anticancer drugs commonly use performance metrics such as power, accuracy of the estimated target dose, or mean absolute error over the estimated dose-response curve (Bornkamp et al., 2007; Dragalin et al., 2010). Equal allocation of subjects to each dose may not fully optimize these metrics. Optimal allocation methods, which include adaptive allocation methods such as the D-optimal (Dragalin et al., 2007), TD-optimal (Dette et al., 2008), and aMCP-Mod (Bornkamp et al., 2011) methods, have been proposed to overcome the limitations of equal allocation. However, they rely on asymptotics, and thus sometimes cannot efficiently optimize the performance metric with the sample size in an actual clinical trial.

Studies of methods in Phase I dose-finding trials for anticancer drugs mainly use the percentages of correct selection (PCS) of MTDs as the performance metric. However, existing methods may not fully optimize the PCS because they do not determine dose escalation to improve PCS.

1.5 Structure of This Thesis and Main Articles

This thesis consists of five chapters. Chapter 2 and 3 are each the main part of one academic article. In Chapter 2, we describe a novel adaptive allocation method using deep reinforcement learning in dose-finding studies for non-anticancer to optimize each performance metrics to be optimized (Matsuura et al., 2022). In Chapter 3, we describe a novel dose escalation method using deep reinforcement learning in dose-finding studies for anticancer drugs to optimize the PCS (Matsuura et al., 2023). Finally, the discussion (Chapter 4) and conclusion (Chapter 5) are presented.

Chapter 2

Optimal Adaptive Allocation in Dose-Response Studies for Non-Anticancer Drugs

2.1 Introduction

Estimation of the dose-response curve for efficacy and selection of the dose for use in confirmatory phase III trials are one of the most difficult decisions in the drug development process. While too low a dose can result in lack of efficacy, too high a dose can cause unnecessary adverse events.

Various methods have been examined to accurately estimate the dose-response curve and ensure correct dose selection. Methods for estimating the dose-response curve include analysis of variance (ANOVA), multiple comparison procedure – modeling (MCP-Mod) method (Bretz et al., 2005), and Bayesian modeling average (BMA)-based method (Ohlssen and Racine, 2015; Gould, 2019). These methods are typically used with equal allocation of subjects for simplicity. Various optimal allocation methods, which include adaptive allocation methods, have been studied (Aouni et al., 2020), such as the D-optimal method (Dragalin et al., 2007), TD-optimal method (Dette et al., 2008), aMCP-Mod method (Bornkamp et al., 2011), and Miller's method (Miller et al., 2007). Studies have evaluated some of these methods (Bornkamp et al., 2007; Dragalin et al., 2010). One common feature in these studies is the evaluation of the operating characteristics in simulation studies using performance metrics, such as statistical power, accuracy of the estimated target dose, and mean absolute error over the estimated dose-response curve.

The issue with equal allocation is that the metrics may not be fully optimized due to non-optimal allocation. Several optimal allocation methods have been proposed in previous studies to overcome this issue, but they rely on asymptotics, and thus sometimes cannot efficiently optimize the performance metric with the sample size in an actual clinical trial. For example, the D-optimal method minimizes the asymptotic variance of the estimates of the dose-response model parameters (Dragalin et al., 2007), and the TD-optimal method minimizes the asymptotic variance of the estimated target dose (Dette et al., 2008).

The purpose of this study is to construct an adaptive allocation rule that can directly optimize the performance metric to be optimized. To achieve this, we use deep reinforcement learning (Sutton and Barto, 2018; Schulman et al., 2017) based on the mean and standard deviation of the response

for each dose and the number of subjects allocated to each dose. A simulation study was conducted to compare the operating characteristics of the equal allocation (commonly used in actual clinical trials), D-optimal method, TD-optimal method, and proposed method.

In Section 2.2, the performance metrics and the proposed method are described. In Section 2.3, the simulation settings and results are presented. In Section 2.4, we summarize and discuss our findings.

2.2 Optimal Adaptive Allocation using Reinforcement Learning

2.2.1 Settings

In most actual dose-response studies, the doses are limited to predetermined discrete values, and thus, we assume this in this study. The number of doses is denoted by K, and the indices of the doses are $k = 1, \ldots, K$, indexed from the lowest dose to the highest dose. The amount of dose is denoted by d_k, where $k = 1$ is the placebo group with $d_1 = 0$. The total number N of subjects to be allocated in a clinical trial is assumed to be predetermined. Each subject is allocated to a dose $k \in \{1, \ldots, K\}$ and response Y is measured. We assume that the clinical team has a performance metric to be optimized, as described in Section 2.2.2, and determines a method to detect dose-response and to estimate the dose-response curve at the end of the trial (e.g., ANOVA, MCP-Mod, or BMA).

In the proposed method, a clinical trial is conducted according to the following steps.

1. At the beginning of the trial, N_{ini} subjects are allocated equally to $k = 1, \ldots, K$ and their responses are obtained.
2. Based on the information obtained so far, each of N_{block} subjects is probabilistically allocated to one of the K doses according to the adaptive allocation rule π^*. Then, their responses are obtained. This step is repeated for $b = 1, \ldots, B$ where $B = (N - N_{\text{ini}})/N_{\text{block}}$.
3. At the end of the trial, the dose-response curve and target dose are estimated, and all performance metrics are evaluated.

In Step 2, the rule π^* selects a dose k so that it can optimize the selected performance metric. The rule π^* is determined before the start of the trial. In Section 2.2.3, we explain how the rule π^* is obtained using deep reinforcement learning.

2.2.2 Performance Metrics

When selecting the target dose to be used in a phase III trial, safety and efficacy of the drug are taken into consideration. For the purpose of simulation studies, we simplify the problem and consider only efficacy for dose selection. In simulation studies, the existence of true dose-response curves is usually assumed to evaluate the methods (Bornkamp et al., 2007; Dragalin et al., 2010). The values of the true and estimated dose-response curves at d_k are denoted by $\mu(d_k)$ and $\hat{\mu}(d_k)$, respectively. To evaluate the operating characteristics of the methods, the following performance metrics are used in general (Bornkamp et al., 2007; Dragalin et al., 2010).

Detecting dose-response

The methods in the previous studies and the proposed method include a decision rule to determine whether the data provides sufficient evidence of dose-response activity. The probability of identifying the presence of dose-response is estimated as the percentage of simulated trials in which the decision rule concluded for dose-response activity. Under a flat dose-response scenario, it gives the type I error rate, and under a non-flat dose-response scenario, it provides the power to make the correct identification of dose-response.

Accuracy of model selection

In several dose-response curve estimation methods, model selection is done from candidate dose-response models such as linear, Emax, and sigmoid Emax models. For the accuracy of the model selection, we calculate the percentage of simulated trials in which the dose-response curve selected in model selection is correct (Mercier et al., 2015), and call this metric "MS". Selecting the correct model is important for estimating the dose-response curve and target dose with small errors.

Accuracy of a target dose

In this study, the target dose d_{targ} is defined as the smallest dose that produces an effect difference from placebo greater than or equal to the clinically relevant target effect δ (minimum effective dose, MED). Here, d_{targ} is a continuous value and is obtained by

$$d_{\text{targ}} = \underset{d\in[d_1,d_K]}{\text{argmin}}\,\{\mu(d) \geq \mu(d_1) + \delta\}.$$

It should be noted that d_{targ} varies with the true dose-response curve. We also consider target effect intervals $I^e_{\text{targ}}(\eta) = \delta(1 \pm \eta)$ (that is, within $\pm 100\eta\%$ of the target effect) and their corresponding target dose intervals $I^d_{\text{targ}}(\eta)$ (Dragalin et al., 2010). The estimated target dose $\hat{d}_{\text{targ}}$ is also a continuous value and is defined using the estimated dose-response curve $\hat{\mu}(d)$ by

$$\hat{d}_{\text{targ}} = \underset{d\in[d_1,d_K]}{\text{argmin}}\,\{\hat{\mu}(d) \geq \hat{\mu}(d_1) + \delta\}.$$

We define the accuracy of the estimated target dose by calculating the percentage of simulated trials in which $\hat{d}_{\text{targ}}$ is correctly within the interval $I^d_{\text{targ}}(0.1)$, and call this metric "TD". In this study, we evaluate "TD" without rounding $\hat{d}_{\text{targ}}$ to the nearest integer because we consider that a better "TD" for the continuous dose also leads to a better "TD" for the discrete dose.

Error in a dose-response curve

Accurate estimation of the dose-response curve is relevant not only for estimating target doses, but also for appropriate labeling after approval. To evaluate the accuracy of the dose-response curve estimation, we calculate the mean absolute error (MAE) between the estimated and true dose-response curves. In actual clinical trials, it is important to determine the effect compared with the placebo group. Therefore, we calculate the MAE after shifting the dose-response curve so that the effect in the placebo group is zero(Dragalin et al., 2010).

$$\text{MAE} = \frac{1}{K-1}\sum_{k=2}^{K} |(\hat{\mu}(d_k) - \hat{\mu}(d_1)) - (\mu(d_k) - \mu(d_1))|$$

2.2.3 Deep Reinforcement Learning

In this section, we describe how to use reinforcement learning (Sutton and Barto, 2018) to obtain an allocation rule that optimizes the selected metric. To conduct reinforcement learning, the distributions for the dose-response curve and observation noise must be given to simulate trials. In each simulated trial, a dose-response curve and responses are probabilistically generated from the distributions. The distributions should reflect the prior beliefs of the clinical team. For example, we can use the candidate models of MCP-Mod with prespecified probabilities when using it to estimate a dose-response curve. Similarly, we can use the prior distributions of BMA when using BMA.

In reinforcement learning, a task is formulated as a Markov decision process (MDP), and an important factor is how to specify the state and reward in the MDP. In the application of an MDP, state s corresponds to a variable that succinctly describes the information available up to that time point. Now, we consider the situation in which the responses of the b-th block have been obtained in Step 2 in Section 2.2.1. In the proposed method, we define s by

$$s = \left\{ \bar{Y}_2 - \bar{Y}_1, \bar{Y}_3 - \bar{Y}_1, \ldots, \bar{Y}_K - \bar{Y}_1, \hat{\sigma}_1, \ldots, \hat{\sigma}_K, \frac{n_1}{N}, \ldots, \frac{n_K}{N} \right\},$$

where $\bar{Y}_k$ and $\hat{\sigma}_k$ are the mean and standard deviation of the responses of the subjects allocated to dose k. The number of subjects allocated to dose k up to that time point is denoted by n_k. Therefore, $\sum_{k=1}^{K} n_k = N$ is satisfied at the end of the clinical trial. s is a vector of the difference from placebo, the standard deviation, and the proportion of the number of subjects allocated.

We define action k to be selected from $\{1, \ldots, K\}$. Unlike when we apply the obtained allocation rule, action k represents that all N_{block} subjects within the b-th block receive the same dose k in the learning. This is to speed up and stabilize reinforcement learning.

Next, we define the reward. For each metric selected from those in Section 2.2.2, we transformed the value into approximately within the range $[0, 1]$ at the end of the trial to use the default value of the learning rate hyperparameter in the software. We write r_x as the reward when the performance metric is x. We define r_{power}, r_{MS}, r_{TD}, and r_{MAE} as follows:

$$r_{\text{power}} = \begin{cases} 1, & \text{if dose-response is detected under a non-flat model} \\ 0, & \text{otherwise} \end{cases}$$

$$r_{\text{MS}} = \begin{cases} 1, & \text{the selected model coincides the true model} \\ 0, & \text{otherwise} \end{cases}$$

$$r_{\text{TD}} = \begin{cases} 1, & \hat{d}_{\text{targ}} \text{ is within the interval } I^d_{\text{targ}}(0.1) \\ 0, & \text{otherwise} \end{cases}$$

$$r_{\text{MAE}} = 1 - 2 \times \text{MAE}.$$

We define $Q_\pi(s, k)$ as the expected cumulative reward from state s by allocating the next block to dose k and after that following the allocation rule π (see Appendix B for the formal definition). The aim of reinforcement learning is to learn the optimal allocation rule π^* such that $\max_k Q_\pi(s, k)$ is maximized for each s. When the number of possible values of s is finite and small, it is possible to use the backward induction method (Lewis and Berry, 1994); however, this method is not feasible in this case. Instead, we express π using a deep neural network (DNN) and obtain π^* numerically by reinforcement learning. Several methods have been proposed to learn π^* (Espeholt et al., 2018;

Fujimoto et al., 2018). Here, we use the proximal policy optimization (PPO) method, a type of deep reinforcement learning, owing to its ease of implementation and high performance (Schulman et al., 2017).

In the PPO method, the probability $\pi(k|s)$ of taking action (in our case, dose) k under state s is represented by a DNN. A DNN with an activation function f and consisting of two intermediate layers with J units can be described as follows:

$$z_j^{(1)} = f\left(\alpha_j^{(1)} + \sum_i \beta_{ji}^{(1)} s_i\right), \qquad z_j^{(2)} = f\left(\alpha_j^{(2)} + \sum_{j'=1}^{J} \beta_{jj'}^{(2)} z_{j'}^{(1)}\right),$$

$$u_k = \alpha_k^{(3)} + \sum_{j'=1}^{J} \beta_{kj'}^{(3)} z_{j'}^{(2)}, \qquad \pi(k) = \mathrm{softmax}(u_k) = \frac{\exp(u_k)}{\sum_{k'=1}^{K} \exp(u_{k'})},$$

where s_i is an element of s, and $\alpha^{(1)}$, $\beta^{(1)}$, $\alpha^{(2)}$, $\beta^{(2)}$, $\alpha^{(3)}$, $\beta^{(3)}$ are the parameters of the DNN.

We estimate π^* using reinforcement learning. Specifically, we first initialize the parameters of the DNN appropriately to initialize π. Then, we simulate a clinical trial according to the current rule π, and obtain the data of the states and rewards. From these data, the parameters of the DNN are updated based on the gradient to increase the reward. We iteratively simulate trials and update them such that π converges to π^*. See Appendix B for the overview of the PPO method.

2.3 Simulation Study

We conducted a simulation study in a slightly modified setting used by Bornkamp et al. (2007) and Dragalin et al. (2010). We compared the performance of the equal allocation, D-optimal method, TD-optimal method, and proposed method.

2.3.1 Design of Simulation Study

We assumed a phase II dose-response study using the MCP-Mod method, which has been used frequently in actual trials in recent years. Note that it is also possible to use reinforcement learning to directly estimate the dose-response curve without using MCP-Mod. Nonetheless, we unified the procedure to use MCP-Mod for a fair comparison with existing methods and to purely evaluate the efficiency of the allocation rules.

In this trial, five doses (0, 2, 4, 6, and 8 mg) were set, and the total sample size was set to 150 subjects. The clinically relevant target effect was $\delta = 1.3$. In MCP-Mod, candidate dose-response models (curves) with the values of their shape parameters must be prepared before the start of the trial. The candidates in this trial were Scenarios 1, 4, and 7 in Table 2.1 with equal probabilities (i.e., 1/3 for each), and the maximum effect in the dose range $[0, 8]$ was assumed to be 1.65. The response was assumed to be the sum of the dose-response curve and the observation noise following a normal distribution with mean 0 and variance 4.5 (Bornkamp et al., 2007). In MCP-Mod, multiple testing with a significance level is performed on the candidates at the end of the trial. The models that pass the testing are fitted to the data, and the shape parameters are estimated. Then, model selection is performed using a predetermined criterion. Here, the significance level was set to 0.025, and model selection was performed using Akaike information criterion (AIC). Finally, the performance metrics in Section 2.2.2 were evaluated using the selected model. Although performance metrics

(except power) are not defined under MCP-Mod in case no model passes the testing, we formally performed model selection using all candidate models and calculated the performance metrics for the evaluation purpose.

For each of the 16 scenarios in Table 2.1, 10,000 simulated trials were used to estimate the mean of the performance metrics. Scenarios 2, 3, 5, 6, 8, and 9 represent the scenarios where the effect was smaller or larger than the candidates, and Scenarios 10 to 15 represent the scenarios where the model was not included in the candidates. These scenarios were set up to verify the robustness of the allocation rule obtained by the proposed method. Scenario 16 was used to evaluate the type I error rate. These scenarios are illustrated in Figure 2.1.

Table 2.1: Dose-response scenarios.

Scenario no.	Model	Max effect	Formula	d_{targ}	$I^d_{\text{targ}}(0.1)$
1	linear	1.65	$\mu(d) = (1.65/8)d$	6.30	(5.67, 6.93)
2	linear	1.65×0.8	$\mu(d) = (1.32/8)d$	7.88	(7.09, 8.00)
3	linear	1.65×1.2	$\mu(d) = (1.98/8)d$	5.25	(4.73, 5.78)
4	Emax	1.65	$\mu(d) = 1.81d/(0.79 + d)$	2.00	(1.44, 2.95)
5	Emax	1.65×0.8	$\mu(d) = 1.45d/(0.79 + d)$	6.83	(3.30, 8.00)
6	Emax	1.65×1.2	$\mu(d) = 2.18d/(0.79 + d)$	1.17	(0.92, 1.52)
7	sigEmax	1.65	$\mu(d) = 1.70d^5/(4^5 + d^5)$	5.06	(4.68, 5.58)
8	sigEmax	1.65×0.8	$\mu(d) = 1.36d^5/(4^5 + d^5)$	7.37	(5.75, 8.00)
9	sigEmax	1.65×1.2	$\mu(d) = 2.04d^5/(4^5 + d^5)$	4.47	(4.24, 4.74)
10	quadratic	1.65	$\mu(d) = (1.65/3)d - (1.65/36)d^2$	3.24	(2.76, 3.81)
11	quadratic	1.65×0.8	$\mu(d) = (1.32/3)d - (1.32/36)d^2$	5.26	(3.98, 8.00)
12	quadratic	1.65×1.2	$\mu(d) = (1.98/3)d - (1.98/36)d^2$	2.48	(2.16, 2.84)
13	exponential	1.65	$\mu(d) = 0.00055(\exp(d) - 1)$	7.76	(7.66, 7.86)
14	exponential	1.65×0.8	$\mu(d) = 0.00044(\exp(d) - 1)$	7.98	(7.88, 8.00)
15	exponential	1.65×1.2	$\mu(d) = 0.00066(\exp(d) - 1)$	7.58	(7.47, 7.67)
16	flat	0	$\mu(d) = 0$	-	-

Note: If the upper of $I^d_{\text{targ}}(0.1)$ did not exist or was greater than 8 (maximum dose), the upper was set to 8.

2.3.2 Allocation Rule

We used the following eight allocation rules: Equal, D-optimal 1, D-optimal 2, TD-optimal 1, TD-optimal 2, RL-power, RL-MS, RL-TD, and RL-MAE. We used the sans-serif font for rule names to distinguish the objective used in RL, which represents reinforcement learning, from the evaluated performance metrics. The details of the eight allocation rules are described below.

Equal

At the beginning of the trial, 150 subjects were equally allocated to five doses ($n_1 = n_2 = n_3 = n_4 = n_5 = 30$). This rule is easy to understand and is most frequently used in actual clinical trials.

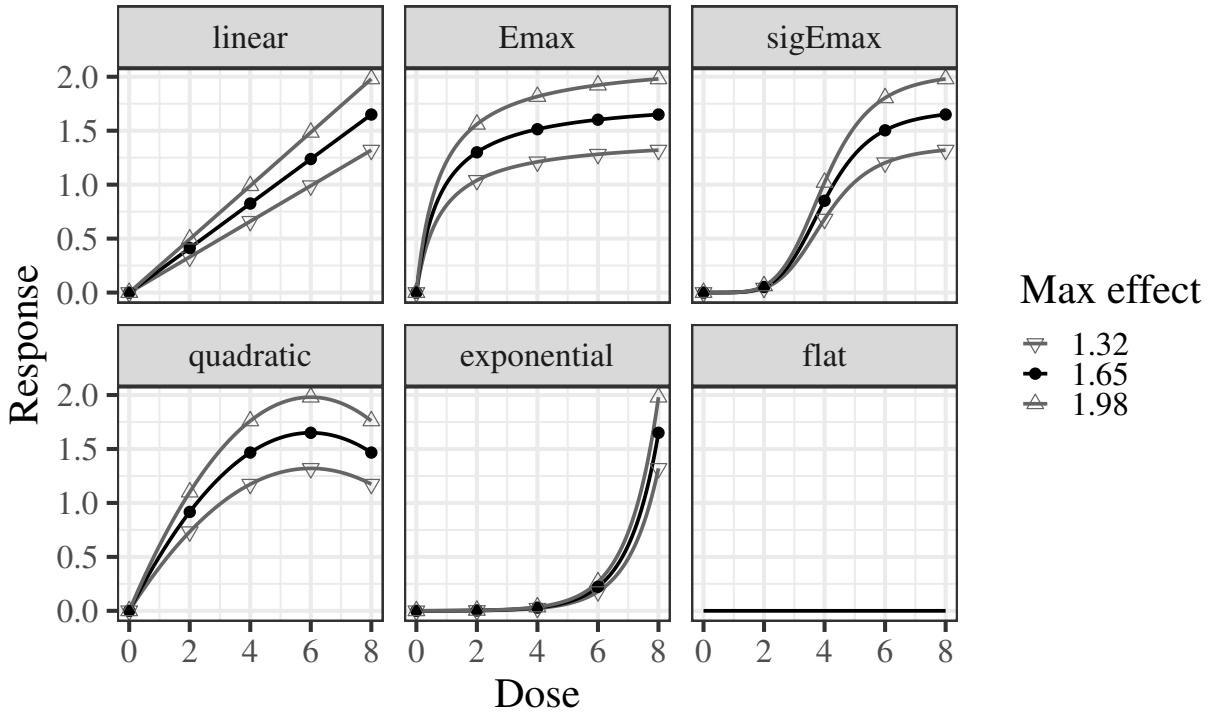

Figure 2.1: Dose-response scenarios.

D-optimal 1

At the beginning of the trial, the allocation ratios were calculated based on the D-optimal method (Dragalin et al., 2007) to minimize

$$-\sum_m \frac{p_m}{k_m} \log\left(\det M_m\right), \tag{2.1}$$

where m is the index of each candidate model, p_m is the prior probability of model m (here, 1/3 for each m), k_m is the number of parameters of model m, and M_m is the Fisher information matrix under model m. The calculated allocation ratios for each group were 0.30, 0.20, 0.12, 0.09, and 0.29, respectively. The calculated ratios were rounded to integer values using the method by Pukelsheim and Rieder (1992), and $n_1 = 44$, $n_2 = 30$, $n_3 = 18$, $n_4 = 14$, $n_5 = 44$ were allocated.

D-optimal 2

The subjects were adaptively allocated based on the D-optimal method (Dragalin et al., 2007). More specifically, at the beginning of the trial, 50 subjects were equally allocated to the five doses. Then, after obtaining their responses, we determined the allocation ratios that minimized Eq. (2.1), given the number of allocated subjects and the number of subjects in the next block (that is, by using the options "nold" and "n" in DoseFinding::optDesign function of R). Here, 10 subjects were allocated in the next block. The model probabilities p_m $(m = 1, 2, 3)$ were set to 1/3 before the trial, and were updated according to Section 5 in Miller et al. (2007) for each block. The shape parameters were not updated and were fixed to those of the candidates (i.e., Scenarios 1, 4, and 7).

The calculated ratios were rounded to integer values using the method by Pukelsheim and Rieder (1992). Then, the responses of the 10 allocated subjects were obtained, and the allocation ratios were calculated again to allocate the next 10 subjects. This was repeated until the total number of subjects reached 150.

TD-optimal 1

At the beginning of the trial, the allocation ratios were calculated based on the TD-optimal method (Dette et al., 2008) to minimize

$$\sum_m p_m \log\left(v_m\right), \tag{2.2}$$

where m is the index of each candidate model, p_m is the probability of model m (here, 1/3 for each m), and v_m is proportional to the asymptotic variance of the estimated target dose under model m. The calculated allocation ratios were 0.31, 0.26, 0.12, 0.18, and 0.14, respectively. According to these ratios, $n_1 = 46$, $n_2 = 39$, $n_3 = 17$, $n_4 = 27$, and $n_5 = 21$ were allocated.

TD-optimal 2

The subjects were adaptively allocated based on the TD-optimal method. The procedure was the same as that used in D-optimal 2, except that the objective function was Eq. (2.2) instead of Eq. (2.1).

RL-power, RL-MS, RL-TD, and RL-MAE

Because the procedures for constructing these rules are similar, RL-MAE is explained as an example.

We simulated clinical trials in reinforcement learning using the settings in Sections 2.2 and 2.3.1, and learned the allocation rule. In each simulated trial, the dose-response curve was determined uniformly at random from the scenarios considered in MCP-Mod (i.e., Scenarios 1, 4, and 7), and the observation noise was generated from a normal distribution with mean 0 and variance 4.5. We used $N_{\text{ini}} = 50$ and $N_{\text{block}} = 10$. In addition, we used ReLU ($f(x) = \max(0, x)$) as the activation function and a DNN consisting of two intermediate layers with 256 units. The settings of the DNN were the default values of the software (Liang et al., 2018). After each simulated trial, the MAE was evaluated. After each 1,000 simulated trials, allocation rule π was updated using the accumulated data of the states and MAEs. With 1,000,000 simulated trials in reinforcement learning, the allocation rule $\pi^*(k|s)$ was obtained. See Appendix B for details on the hyperparameters of the PPO method.

At the beginning of the trial, 50 subjects were allocated equally to the five doses. Thereafter, each time the responses were obtained, each of the 10 subjects was probabilistically allocated to one of the five doses according to the discrete distribution $\pi^*(k|s)$. This was repeated until the total number of subjects reached 150.

RL-power, RL-MS, and RL-TD, were the same as RL-MAE, except that the metrics to be optimized were power, MS, and TD in Section 2.2.2.

In general, it is known that using p-values without considering adaptive allocation may inflate the type I error rate, and a simulation-based method to control the type I error rate has been

discussed previously (Bretz et al., 2008; FDA, 2017). Here, we first calculated the p-values for the flat scenario, and then adjusted the significance level threshold based on the distribution of the p-values. Then, using the adjusted significance level, we simulated the other scenarios and evaluated the performance metrics.

For deep reinforcement learning, we used the RLlib library in Python (Liang et al., 2018) and for the MCP-Mod, D-optimal, and TD-optimal methods, we used the DoseFinding package in R (Bornkamp et al., 2009). The code with hyperparameters is available in our GitHub repository (Matsuura, 2022), which can be modified according to the requirement.

2.3.3 Results

In this section, the means of the performance metrics obtained from 10,000 simulations for each allocation rule are presented.

The results for the type I error rate are shown in Figure 2.2. Figure 2.2 (a) shows the type I error rate of each rule when the significance level was 0.025. Note that this significance level was based on MCP-Mod, and the type I error rate was not theoretically guaranteed for adaptive allocation rules. Figure 2.2 (b) shows the type I error rates of the proposed methods using various significance levels. From these results, we adjusted the significance level to 0.0235 for RL-power, 0.024 for RL-MS, 0.021 for RL-TD, and 0.0165 for RL-MAE to control the type I error rate. We continued to use a significance level of 0.025 for the other rules, assuming that fluctuations around the 2.5% level were consistent with the Monte Carlo error and the type I error rates were under control. Using these adjusted significance levels, we evaluated the performance metrics for the other scenarios.

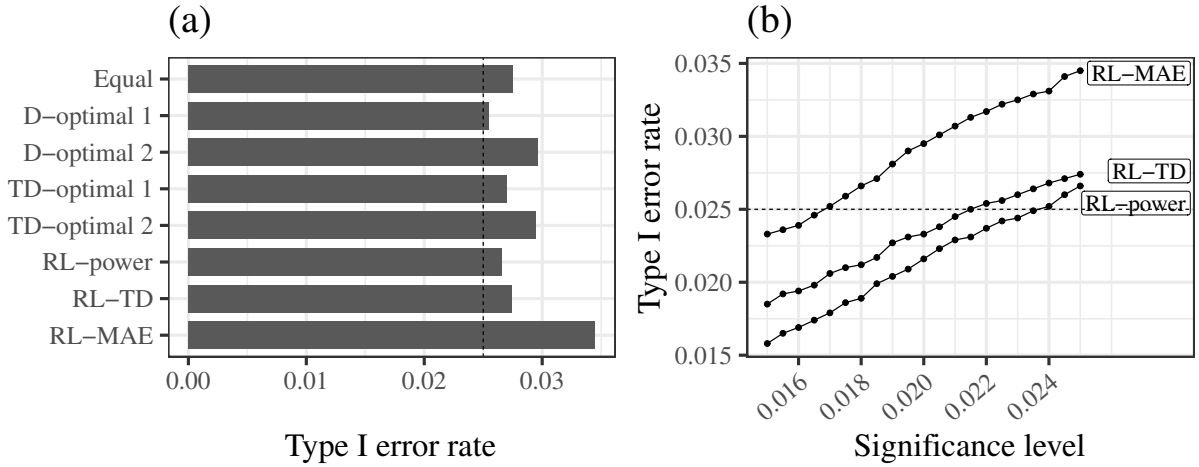

Figure 2.2: The results for the type I error rate before adjustment.

The results of the performance metrics (that is, power, MS, TD, and MAE) were similar for the four models (linear, Emax, sigEmax, and quadratic), whereas the results were different for the exponential model. Here, the average results over the all models are shown. For the results of each model, see Figures A.1–A.4 in Appendix A.

The results for power are shown in Figure 2.3. RL-power certainly improved power. In contrast, RL-MS and RL-MAE worsened the power. The lower average power of RL-MAE may be due to the much lower power when the exponential model was true (see Appendix A). Notably, RL-power had

high power even when the true maximum effect was smaller and larger than the candidates, even though RL-power was trained assuming a maximum effect of 1.65.

The results for MS, TD, and MAE (Figure 2.4–2.6) were calculated from simulations where multiple testing was significant. We confirmed that the results were almost the same, even if we included the simulations in which the testing was not significant. The results for the MS are shown in Figure 2.4. RL-MS certainly improved the MS. In contrast, RL-power worsened the MS. RL-MS was also effective in scenarios different from the candidates. The results for the TD are shown in Figure 2.5. Better results were obtained when the maximum effect was smaller than that of the candidates. This may be because of the wider range of $I^d_{\text{targ}}(0.1)$. RL-TD and RL-MAE improved the TD. Note that these rules were better than TD-optimal 1 and 2. In contrast, RL-power worsened the TD. The results for the MAE are shown in Figure 2.6. RL-MAE improved the MAE. In contrast, RL-MS worsened the MAE. RL-MAE was also effective in scenarios that were different from the candidates.

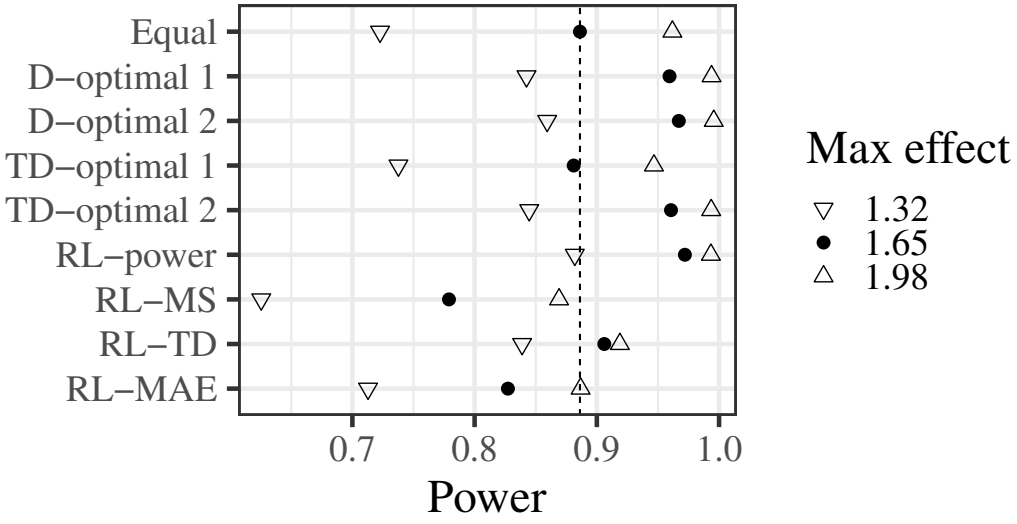

Figure 2.3: The results for power. The vertical dotted line represents the value of the equal allocation.

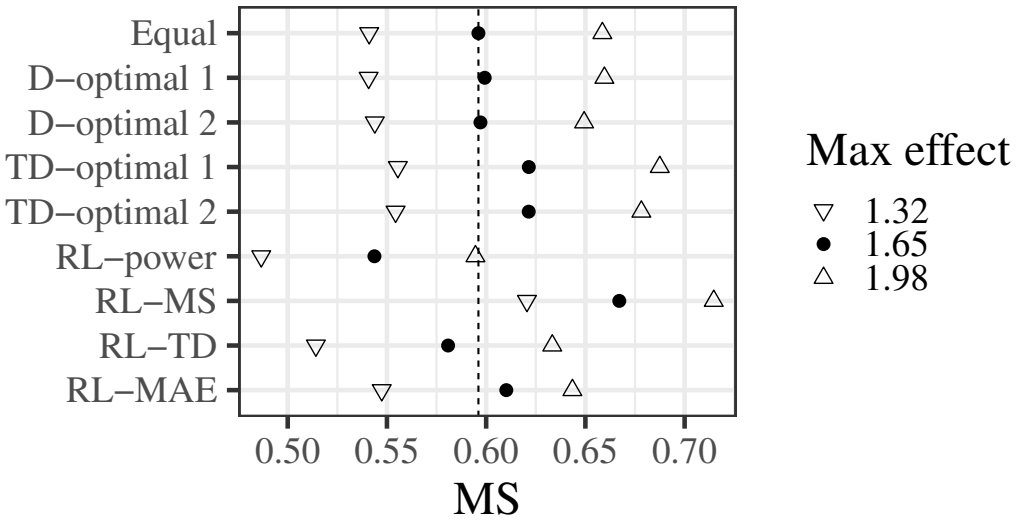

Figure 2.4: Probability of selecting the true model.

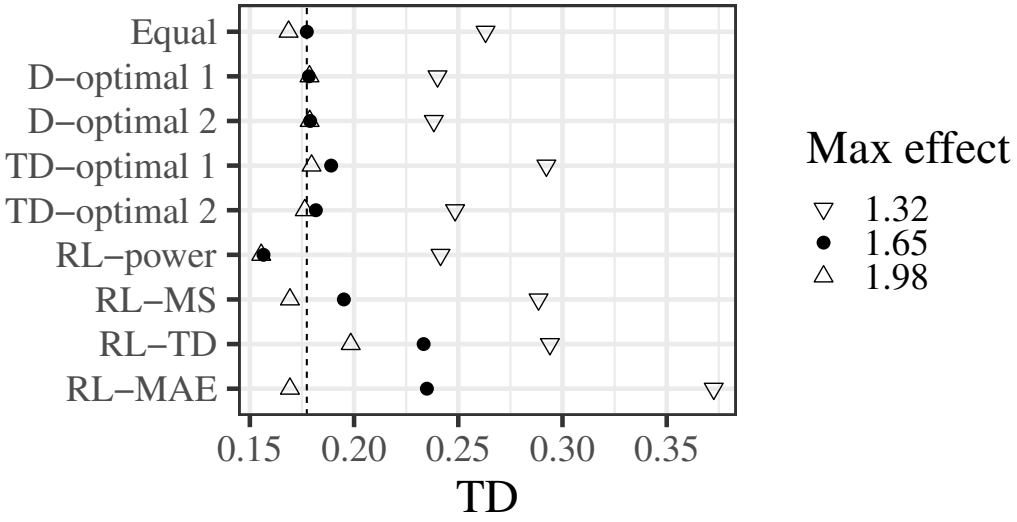

Figure 2.5: Probability that the estimated target dose is within the interval $I^d_{\text{targ}}(0.1)$.

In summary, these results showed that the proposed method improves not only the performance metric used for optimization, but also many other metrics. In particular, RL-MAE was superior in most metrics for correctly estimating the dose-response relationship for phase III trials.

The average number of subjects allocated to each dose is shown in Figure 2.7. This figure shows that the proposed methods tended to allocate more subjects to 0 mg than Equal. In addition, RL-MS and RL-MAE tended to allocate more subjects to 2 mg. Since the allocation that optimizes power for the contrast test (assuming the same variance across the dose groups) should be the allocation that places half of the subjects on placebo and the other half on the dose providing the maximum effect, it is natural that RL-power tended to allocate more subjects to 0 and 8 mg. Since 0, 2, and 8 mg are likely to be important in distinguishing the flat and Emax models from the rest, it is natural that

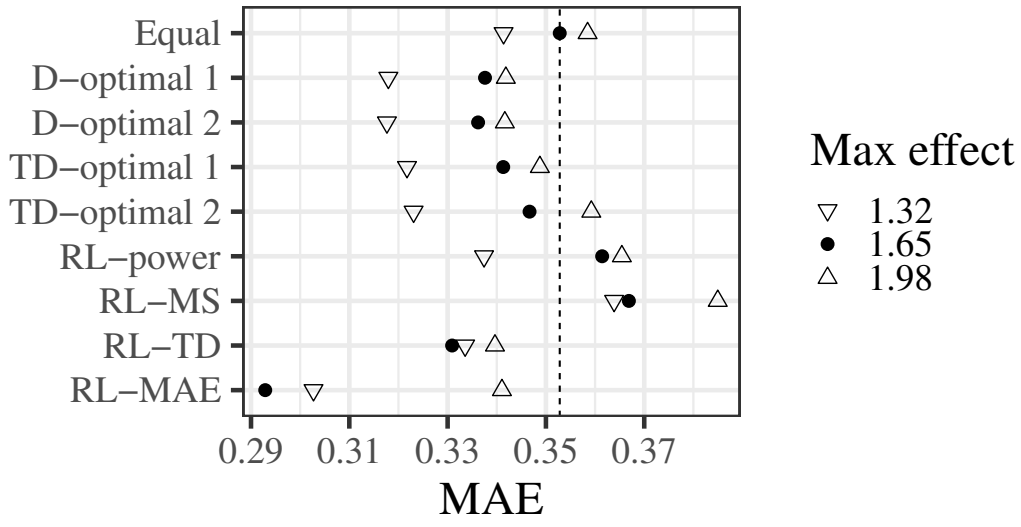

Figure 2.6: The results for MAE. Smaller MAE implies better accuracy.

more subjects will be allocated to these doses. For the results of each model, see Figures A.5–A.7 in Appendix A.

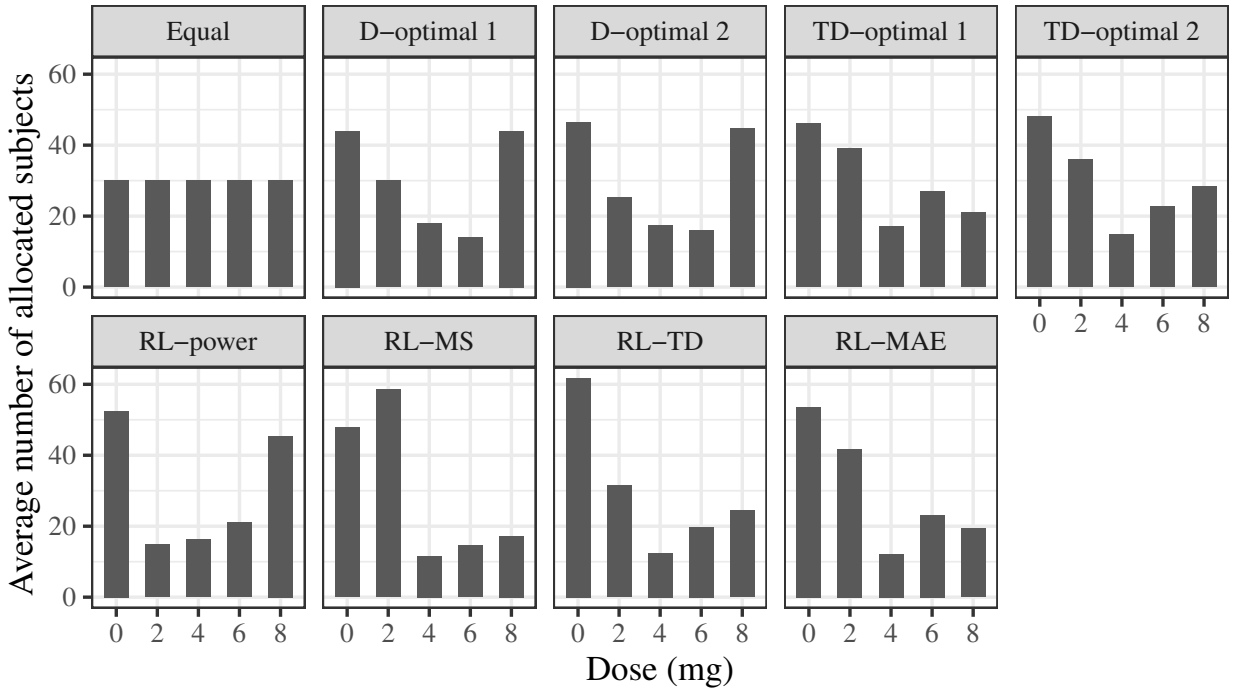

Figure 2.7: The results for the average number of subjects allocated.

Note that the good performance of RL-MAE was not only due to the non-uniform allocation, but also from the adaptivity of the allocation. In fact, we confirmed that the performance does not improve if we use a fixed design with the number of subjects equal to the average of those of RL-MAE in Figure 2.7. See Appendix A for details.

2.4 Summary

We showed that deep reinforcement learning with an appropriately defined state and reward can be used to construct adaptive allocation rules that can directly optimize the performance metrics to be optimized. In general, reinforcement learning becomes difficult when the reward (i.e., the performance metric evaluated at the end of each trial) is delayed, and the observation is noisy. Phase II trials have these difficulties, and it is not obvious whether reinforcement learning works successfully to address the same. Nonetheless, we have shown that it can work well if we appropriately design and choose the Markov decision process as well as the learning algorithm and hyperparameters.

Chapter 3

Optimal Dose Escalation Methods in Dose-Finding Studies for Anticancer drugs

3.1 Introduction

One of the most important objectives in phase I trials of a novel anticancer drug is to identify the maximum tolerated dose (MTD) (Green et al., 2012). The MTD is typically defined as the dose at which the probability of occurrence of a dose-limiting toxicity (DLT) is closest to the target DLT probability, which is a predetermined acceptable probability of occurrence. Because phase I oncology trials usually include subjects with advanced cancer who did not respond to standard therapies, the number of subjects included in the trial is low. Therefore, it is necessary to identify the MTD in a limited number (approximately a few dozen) of subjects.

Many designs have been proposed to determine the MTD. These designs can be classified into rule-based, model-based, and model-assisted designs (Le Tourneau et al., 2009; Daimon et al., 2019). In rule-based design, rules for dose escalation are determined before the trial, and the standard 3+3 design is often used. In model-based design, the relationship between dose and DLT probability is represented by a statistical model, and the continual reassessment method (CRM) proposed by O'Quigley et al. (1990) is often used. In model-assisted design, rules for dose escalation are determined using a statistical model before the trial, and the BOIN method proposed by Liu and Yuan (2015) and the Keyboard method proposed by Yan et al. (2017) are often used. Several studies have evaluated the operating characteristics of these methods (Zhou et al., 2018b, 2018a; Horton et al., 2017). Zhou et al. concluded that the CRM and BOIN methods have similar operating characteristics, and that the percentages of correct selection (PCS) of the MTDs are also similar.

However, the problem is the low PCS from each of the methods listed above in scenarios that would be expected in actual trials. For example, the average PCS for the eight representative scenarios listed in Appendix C of Zhou et al. (2018b) was less than 50%. Determining an overly toxic dose as the MTD will cause ethically unacceptable DLTs in late phase trials. Determining a dose that is too safe as the MTD will lead to a lack of efficacy and failure of late phase trials. It is important to develop a method to improve the PCS as much as possible in order to succeed in late phase trials.

The objective of this study is to construct an action selection rule with a higher PCS than the

above methods. The action options are as follows: escalate the dose, de-escalate the dose, stay at the same dose, or stop the trial and determine the MTD. We achieve this purpose by deep reinforcement learning (Sutton and Barto, 2018; Schulman et al., 2017) based on the number of subjects and DLTs at each dose. A simulation study was conducted to evaluate the operating characteristics of the proposed method and to compare the PCS with those of the 3+3 design, CRM, BLRM (Neuenschwander et al., 2008), BOIN, mTPI (Ji et al., 2010), and i3+3 (Liu et al., 2020) methods.

In Section 3.2, the proposed method is described. In Section 3.3, the simulation settings and results are presented. In Section 3.4, we summarize and discuss our findings.

3.2 Method

3.2.1 Settings

We assume that the doses are limited to a predetermined number of J discrete values and denote each dose level by j. The target DLT probability is denoted by ϕ. The DLT probability is assumed to increase monotonically with the dose level, and the DLT probability at dose level j is denoted by p_j. In this study, we call the combination of $(p_1, \ldots, p_J)$ a scenario. The maximum sample size N is assumed to be predetermined. A trial consists of several cohorts, and one dose is selected for each cohort. A cohort contains one to five subjects, and a cohort of three subjects is often used in actual trials.

In the proposed method, clinical trials are conducted according to the following steps:

1. Assign subjects in the first cohort to the lowest dose and evaluate their DLTs.
2. Based on the action selection rule π^* and the information obtained so far, select the dose for the next cohort and continue the trial, or determine the MTD (including no MTD) and stop the trial. Evaluate DLTs if the trial continues. This step is repeated until the maximum sample size is reached or the action option of stopping the trial is selected.

In Step 2, the rule π^* selects an action so that it can optimize the PCS at the end of the trial. Note that the rule π^* is determined before the start of the trial. In Section 3.2.2, we explain how to obtain the rule π^* using deep reinforcement learning.

The obtained dose-escalation rule π^* sometimes becomes too aggressive because it is designed to purely optimize the PCS. For such cases, ad-hoc rules can be added for overdose control. An example will be presented later in the simulation study.

3.2.2 Deep Reinforcement Learning

In this section, we describe how to use reinforcement learning (Sutton and Barto, 2018) to obtain an action selection rule that optimizes the PCS. In our method, the acceptable range of target DLT probabilities is defined as $[\phi - \varepsilon, \phi + \varepsilon]$. The unacceptable range of target DLT probabilities is defined as $[0, \phi - \delta]$ and $[\phi + \delta, 1]$. Here, $\delta > \varepsilon$ are predetermined parameters.

To conduct reinforcement learning, the distribution of the scenarios must be determined to simulate trials. In each simulated trial, a scenario is probabilistically generated from the distribution, and

the DLTs are generated according to the scenario. In general, it is important to generate scenarios in which the MTD is clearly defined in order to stably learn π^*. Thus, we propose to generate scenarios by the following procedure based on J, ϕ, ε, and δ. We use a logistic model for all scenarios: $\text{logit}(p_j) = a + b \times j$ for dose levels $j = 1, \ldots, J$.

- $J+1$ scenarios with index $i = 0, \ldots, J$ for middle p_i: a and b are determined such that $p_i = \phi$ and $p_{i+1} = \phi + \delta$. $i = 0$ is the case where there is no MTD.
- $J+1$ scenarios with index $i = 0, \ldots, J$ for lower p_i: a and b are determined such that $p_i = \phi - \varepsilon$ and $p_{i+1} = \phi + \delta$.
- J scenarios with index $i = 1, \ldots, J$ for higher p_i: a and b are determined such that $p_{i-1} = \phi - \delta$ and $p_i = \phi + \varepsilon$.

In each of these $3J + 2$ scenarios, dose i is clearly the MTD. For each dose level, p_j close to 0.0 or 1.0 may be a little unrealistic, and so we clip the value to 0.05 when $p_j < 0.05$, and to 0.8 when $p_j > 0.8$.

In reinforcement learning, a task is formulated as a Markov decision process (MDP), and the important factor is how to specify the state, action, and reward in the MDP. In the application of an MDP, state s corresponds to a variable that succinctly describes the information available up to that time point. Now, consider the situation where the dose level used in the previous cohort is j', the cumulative number of subjects assigned to each dose is n_j $(j = 1, \ldots, J)$, and the cumulative number of DLTs at each dose is x_j. In the proposed method, we define s by

$$s = \left\{ \frac{j'}{J}, \frac{n_1}{N}, \frac{n_2}{N}, \ldots, \frac{n_J}{N}, \frac{x_1}{N}, \ldots, \frac{x_J}{N}, \frac{\sum_{j=1}^{J} n_j}{N}, \frac{\sum_{j=1}^{J} x_j}{N} \right\}.$$

s is a vector of the previous dose level, the proportion of the number of subjects assigned, and the proportion of the number of DLTs.

We define action a to be selected from the following options:

$$a \in \{d_{\text{de-escalate}}, d_{\text{stay}}, d_{\text{escalate}}, \text{stop}_1, \text{stop}_2, \ldots, \text{stop}_J\},$$

where $d_{\text{de-escalate}}$, d_{stay}, and d_{escalate} are the actions of selecting the dose for the next cohort and continuing the trial, representing de-escalating the dose to level $j' - 1$, staying at the current dose level j', and escalating the dose to level $j' + 1$, respectively. If $d_{\text{de-escalate}}$ is selected when $j' = 1$ (i.e., the lowest dose), we determine no MTD and stop the trial. If d_{escalate} is selected when $j' = J$ (i.e., the highest dose), we decide to use the highest dose for the next cohort. The $\text{stop}_1, \ldots, \text{stop}_J$ are the actions to determine the MTD and stop the trial, where stop_j determines the dose level j as the MTD.

We define reward r obtained at the end of the trial as $r = \text{I(MTD)}$, where I(MTD) is an indicator function, which is equal to 1 if the determined dose is consistent with the MTD, and 0 otherwise. If the total number of subjects reached N but an action to continue the trial is selected, the trial is stopped immediately, and the reward is defined as 0.

We define $Q^\pi(s, a)$ as the expected reward from state s by selecting action a and after that following the action selection rule π (see Appendix B for the formal definition). The aim of reinforcement learning is to learn the optimal rule π^* such that $\max_a Q^\pi(s, a)$ is maximized for each s. When the number of possible values of s is finite and small, it is possible to use the backward induction method

(Lewis and Berry, 1994). However, this method is not feasible in this case. Instead, we express π using a deep neural network (DNN) and obtain π^* numerically by reinforcement learning. Several methods have been proposed to learn π^* (Espeholt et al., 2018; Fujimoto et al., 2018). Here, we use the proximal policy optimization (PPO) method, a type of deep reinforcement learning, owing to its ease of implementation and high performance (Schulman et al., 2015).

In the PPO method, the probability $\pi(a|s)$ of selecting action a under state s is represented by a DNN. A DNN with an activation function f and consisting of two intermediate layers with L units can be described as

$$z_l^{(1)} = f\left(\xi_l^{(1)} + \sum_i \eta_{li}^{(1)} s_i\right), \qquad z_l^{(2)} = f\left(\xi_l^{(2)} + \sum_{l'=1}^{L} \eta_{ll'}^{(2)} z_{l'}^{(1)}\right),$$

$$u_a = \xi_a^{(3)} + \sum_{l'=1}^{L} \eta_{al'}^{(3)} z_{l'}^{(2)}, \qquad \pi(a) = \mathrm{softmax}(u_a) = \frac{\exp(u_a)}{\sum_{a'} \exp(u_{a'})},$$

where s_i is an element of s, and $\xi^{(1)}$, $\eta^{(1)}$, $\xi^{(2)}$, $\eta^{(2)}$, $\xi^{(3)}$, $\eta^{(3)}$ are the parameters of the DNN.

We estimate π^* using reinforcement learning. To be more specific, we first initialize the parameters of the DNN appropriately to initialize π. Then, we simulate a clinical trial according to the current rule π, and obtain the data of the states, actions, and rewards. From these data, the parameters of the DNN are updated based on the gradient to increase the reward. We iteratively simulate trials and update them so that π converges to π^* (Schulman et al., 2015). When we learn π^* in reinforcement learning, action a is probabilistically selected according to a discrete probability distribution $\pi(a|s)$, while when we apply the learned rule to a clinical trial, action a is determined by $a^* = \underset{a}{\mathrm{argmax}}\, \pi^*(a|s)$ in our method.

3.3 Simulation Study

We conducted a simulation study to compare the performance of the 3+3 design, CRM, BLRM, BOIN, mTPI, i3+3, and the proposed methods—Reinforcement Learning based Escalation (RLE), and RLE with ad-hoc rules (modified RLE; mRLE).

3.3.1 Simulation Settings

The simulation study was conducted under the same settings as in Appendix C in Zhou et al. (2018b). Thus, the number of dose levels was $J = 6$, the maximum sample size was $N = 36$, and the cohort size was three. The target DLT probability was $\phi = 0.25$, and the 10 scenarios shown in Table 3.1 were used to evaluate the operating characteristics of the methods. The MTDs for each scenario are shown in bold. The first eight scenarios were selected by Zhou et al. as the representative scenarios from 1,000 randomly generated scenarios using the method of Clertant and O'Quigley (2017) with $\phi = 0.25$. We added the last two scenarios, which were not similar to the hypothetical scenarios used in reinforcement learning. In Scenario 7, all dose levels were overly toxic, so there was no MTD. For each of the 10 scenarios, 10,000 simulated trials were conducted to estimate the mean of the performance metrics described in Section 3.3.2.

For the CRM, BLRM, BOIN, and mTPI methods, we used the same settings as those used by Zhou et al. (2018b). Specifically, in CRM, we used the one-parameter power model given by

$$p_j = \alpha_j^{\exp(\alpha)} \text{ for } j = 1, \ldots, J,$$

Table 3.1: The 10 scenarios to evaluate the operating characteristics.

	Dose level					
Scenario no.	1	2	3	4	5	6
1	**0.26**	0.34	0.47	0.64	0.66	0.77
2	0.18	**0.25**	0.32	0.36	0.60	0.69
3	0.09	0.16	**0.23**	0.34	0.51	0.74
4	0.07	0.12	0.17	**0.27**	0.34	0.55
5	0.03	0.13	0.17	0.19	**0.26**	0.31
6	0.04	0.05	0.09	0.14	0.15	**0.24**
7	0.34	0.42	0.46	0.49	0.58	0.62
8	**0.13**	0.41	0.45	0.58	0.75	0.76
9	0.05	0.08	0.11	**0.15**	0.60	0.72
10	0.15	0.17	0.19	0.21	0.23	**0.25**

where α is a parameter, and $0 < \alpha_1 < \ldots < \alpha_J < 1$ are the prior guess for the DLT probability at each dose, which is often called the "skeleton" of CRM. Here, the skeleton was set to $(\alpha_1, \ldots, \alpha_J) = (0.062, 0.140, 0.25, 0.376, 0.502, 0.615)$. For the BLRM method, the dosages of six doses were set to $(12.5, 25, 50, 100, 150, 200)$ mg, and the reference dose was $d^* = 200$ mg. We used the vague bivariate normal distribution for the prior of $(\log \alpha, \log \beta)$ such that

$$\begin{pmatrix} \log \alpha \\ \log \beta \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} -0.837 \\ 0.381 \end{pmatrix}, \begin{pmatrix} 2.015^2 & 0 \\ 0 & 1.027^2 \end{pmatrix} \right).$$

The proper dosing interval was set to $(0.2, 0.3)$. For the BOIN method, we followed the guidance of Liu and Yuan (2015) and used $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$, where ϕ_1 is the highest DLT probability that is deemed to be underdosing such that dose escalation is needed; ϕ_2 is the lowest DLT probability that is deemed to be overdosing such that dose de-escalation is needed. For the mTPI and i3+3 methods, the proper dosing interval was set to $(0.2, 0.3)$. In the above methods, the trial was stopped, and no MTD was determined if $\Pr(p_1 > \phi \,|\, \text{data}) > 0.95$.

In the proposed method, $\varepsilon = 0.04$ and $\delta = 0.1$ were set. We simulated clinical trials in reinforcement learning using the settings in Sections 3.2.2 and learned the action selection rule. In each simulated trial, a scenario was generated uniformly at random from the 20 scenarios in Table 3.2, and the DLTs were generated according to the scenario. After each simulated trial, the correctness of the dose determined as the MTD was evaluated, and the action selection rule π was updated.

We used ReLU ($f(x) = \max(0, x)$)) as the activation function and a DNN consisting of two intermediate layers with 256 units (Figure 3.1). The settings of the DNN were the default values of the software (Liang et al., 2018). With approximately 3,000,000 simulated trials in reinforcement learning, the allocation rule $\pi^*(a|s)$ was obtained. RLE used this rule $\pi^*(a|s)$. In addition to the rule $\pi^*(a|s)$, mRLE added the ad-hoc rules that prohibited escalate if $x_{j'}/n_{j'} > \phi$, stay and escalate if $x_{j'}/n_{j'} > 2\phi$, and de-escalate if $x_{j'}/n_{j'} < \phi$. For RLE and mRLE, we ran 10,000 simulations for each of the 10 scenarios to evaluate the operating characteristics, as shown in Table 3.1.

For deep reinforcement learning, we used the RLlib library in Python (Liang et al., 2018); and for the 3+3 design, CRM, BOIN, and mTPI methods, we used the Escalation package in R by Kristian Brock; and for the BLRM method, we used the rstan package in R (Carpenter et al., 2017). The

Table 3.2: The 20 scenarios used in reinforcement learning.

	Dose level					
Scenario no.	1	2	3	4	5	6
1	**0.25**	0.35	0.465	0.584	0.694	0.786
2	0.171	**0.25**	0.35	0.465	0.584	0.694
3	0.113	0.171	**0.25**	0.35	0.465	0.584
4	0.073	0.113	0.171	**0.25**	0.35	0.465
5	0.05	0.073	0.113	0.171	**0.25**	0.35
6	0.05	0.05	0.073	0.113	0.171	**0.25**
7	0.35	0.465	0.584	0.694	0.786	0.8
8	**0.21**	0.35	0.522	0.688	0.8	0.8
9	0.116	**0.21**	0.35	0.522	0.688	0.8
10	0.061	0.116	**0.21**	0.35	0.522	0.688
11	0.05	0.061	0.116	**0.21**	0.35	0.522
12	0.05	0.05	0.061	0.116	**0.21**	0.35
13	0.05	0.05	0.05	0.061	0.116	**0.21**
14	0.35	0.522	0.688	0.8	0.8	0.8
15	**0.29**	0.486	0.686	0.8	0.8	0.8
16	0.15	**0.29**	0.486	0.686	0.8	0.8
17	0.071	0.15	**0.29**	0.486	0.686	0.8
18	0.05	0.071	0.15	**0.29**	0.486	0.686
19	0.05	0.05	0.071	0.15	**0.29**	0.486
20	0.05	0.05	0.05	0.071	0.15	**0.29**

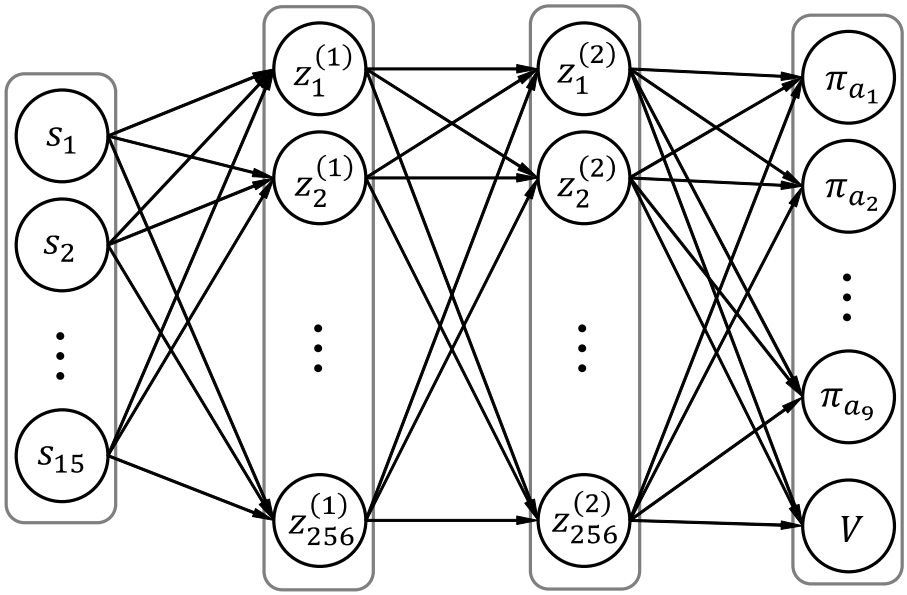

Figure 3.1: Architecture plot of the deep neural network (DNN). For V in the last layer, see Appendix B.

code with hyperparameters is available in our GitHub repository (Matsuura, 2023), which can be modified according to the requirements.

3.3.2 Performance Metrics

The following 10 performance metrics from (a)–(j) were evaluated.

(a) Percentage of correct selection (PCS) of the MTD: This is defined as the percentage of simulated trials in which the correct dose was determined as the MTD. In Scenario 7, we evaluated the percentage of no MTD determination because there was no MTD.

(b) Percentage of overdose selection: This is defined as the percentage of simulated trials in which a higher dose than the MTD was selected. Determining an overdose as the MTD carries the risk of causing ethically unacceptable DLTs in late phase trials. Therefore, it is desirable for this percentage to be smaller.

(c) Percentage of underdose selection: This is defined as the percentage of simulated trials in which a lower dose than the MTD was selected. Determining an underdose as the MTD carries the risk of leading to a lack of efficacy and failure of late phase trials. Therefore, it is desirable for this percentage to be smaller.

(d) Number of subjects treated: This is defined as the average number of subjects who were enrolled in a trial.

(e) Number of DLTs: This is defined as the average number of DLTs in a trial. From an ethical viewpoint, it is desirable to have a smaller number of DLTs during a trial.

(f) Number of subjects treated at the MTD: This is defined as the average number of subjects who were assigned to the MTD in a trial. It is desirable to have a higher number of these subjects to increase the number of subjects who receive appropriate treatment.

(g) Number of subjects treated at an overdose: This is defined as the average number of subjects who were assigned to an overdose in a trial. From an ethical viewpoint, it is desirable to have a smaller number of these subjects.

(h) Number of subjects treated at an underdose: This is defined as the average number of subjects who were assigned to an underdose in a trial. It is desirable to have a smaller number of these subjects.

(i) Percentage of incoherent escalation: This is defined as the percentage of clinical trials in which escalation was selected if $x_{j'}/n_{j'} > \phi$ at least once during a trial. It is desirable for this percentage to be smaller because in practice trial physicians or safe monitoring committees usually do not allow such a decision.

(j) Percentage of incoherent de-escalation: This is defined as the percentage of clinical trials in which de-escalation was selected if $x_{j'}/n_{j'} < \phi$ at least once during a trial. It is desirable for this percentage to be smaller.

3.3.3 Results

This section shows the means of each performance metric in Section 3.3.2, based on 10,000 simulations for each scenario. Here, "average" represents the average over the 10 representative scenarios. The averages were calculated after excluding scenarios where the percentage must be zero. For the 3+3 design, CRM, BLRM, BOIN, and mTPI methods, the results roughly reproduced those of Appendix C in Zhou et al. (2018b).

The results for performance metrics (a)–(c) are shown in Figure 3.2. The proposed method, RLE, outperformed other methods in many scenarios (i.e., Scenarios 3, 4, 5, 6, and 10), and improved the PCS. In Scenarios 1 and 2, the PCS of RLE slightly worsened. Although Scenarios 8, 9, and 10 were not similar to the hypothetical scenarios used in reinforcement learning, the PCS of RLE was comparable in Scenario 8 and 9, and better in Scenario 10. For each scenario, RLE stably performed better in the scenario where the MTD was on the high-dose level. This is probably because it is easier to collect enough information from the results of subjects who were assigned to the low dose levels, and thus the learning results from reinforcement learning can be used more efficiently. In actual trials, RLE has an advantage because the clinical team often prepares several low-doses, which are assumed to be safer. The results for (b) show that an overdose was often selected with BLRM in many scenarios, BOIN in Scenario 9, and RLE in Scenarios 2 and 5. The results for (c) show that the 3+3 design, CRM, BOIN, mTPI, and i3+3 had high percentages of underdose selection, indicating the dose was not sufficiently escalated for late phase trials. Although mRLE had generally similar results to RLE, the PCS (a) was lower, and the percentage of overdose selection (b) was higher in Scenarios 3–5. The reason may be that the reduced frequency of aggressively trying higher doses caused the larger estimation error of the MTD.

The results for performance metric (d) are shown in Figure 3.3. The average sample size was almost the same for all methods except for the 3+3 design.

The results for performance metric (e) are shown in Figure 3.4. RLE had more DLTs than the other methods by an average of 2.5 DLTs in the trial with 36 subjects. mRLE had a similar number of DLTs to BLRM.

The results for performance metrics (f)–(h) are shown in Figure 3.5. The results for (f) show that BOIN and RLE, on average, treated fewer subjects at the MTD than other methods. The results for (g) and (h) show that RLE tended to treat more subjects at an overdose, while the other methods tended to treat more subjects at an underdose. RLE often tried an overdose that was one level higher than the MTD in order to escalate the dose sufficiently.

The results for performance metrics (i) and (j) are shown in Figure 3.6. CRM, BLRM, and RLE had high percentages of incoherent escalation, while the other methods had low percentages.

3.3.4 Examples of Dose Escalation Path

When we apply the learned rule in our method to a clinical trial, once state s is determined, the selected action is uniquely determined by $a^* = \underset{a}{\operatorname{argmax}}\, \pi^*(a|s)$. Examples of the dose escalation path for a single simulated trial are shown in Table 3. Cases 1 and 2 show the first five cohorts for RLE and mRLE, respectively.

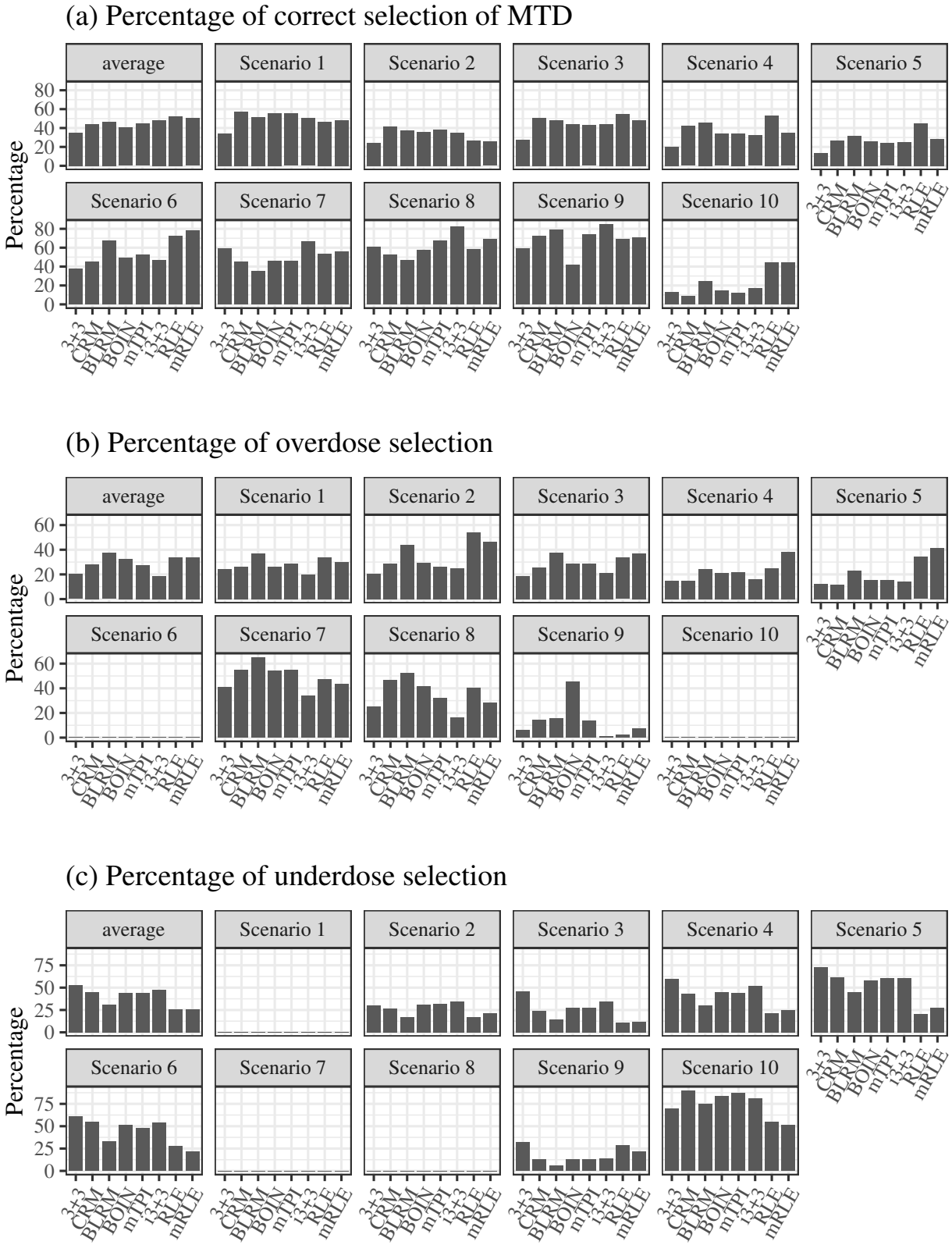

Figure 3.2: The results for performance metrics (a)–(c).

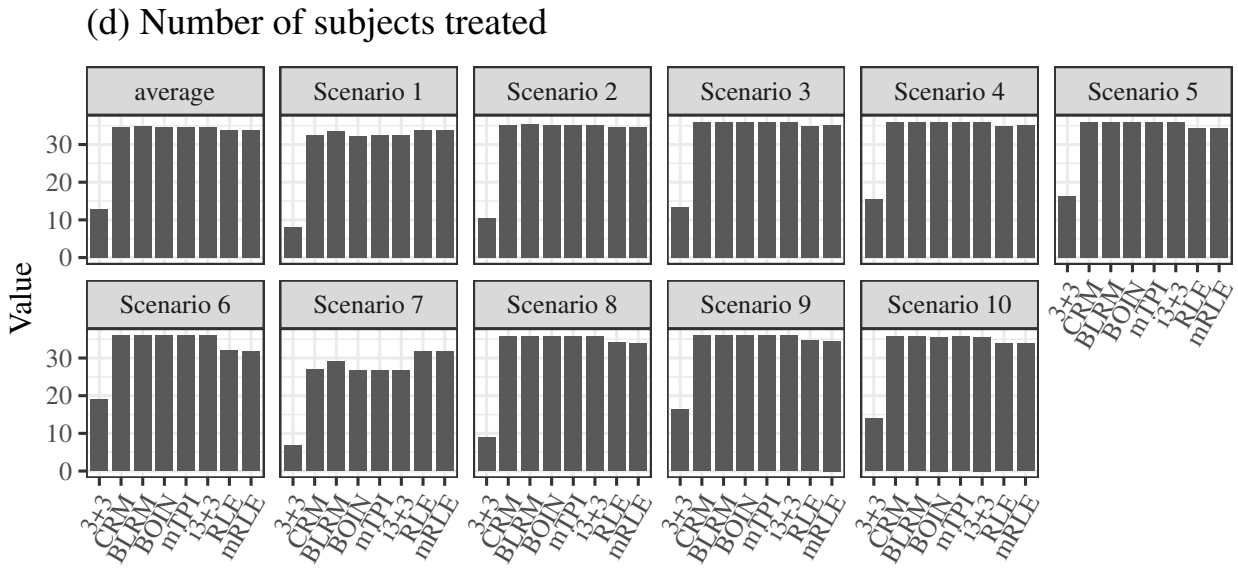

Figure 3.3: The results for performance metric (d).

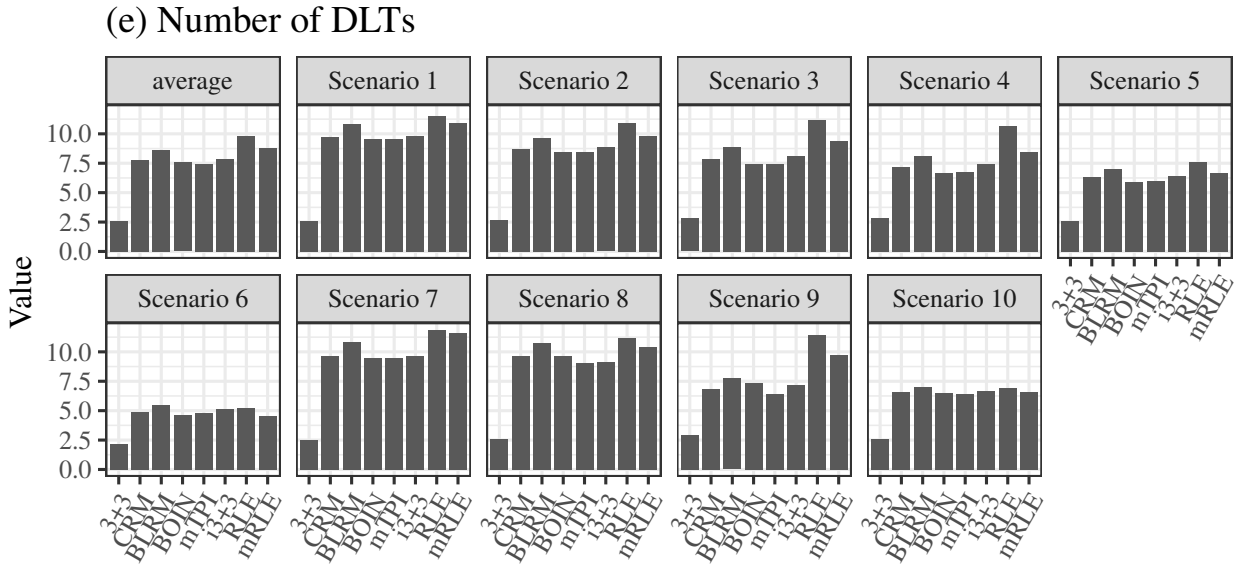

Figure 3.4: The results for performance metric (e).

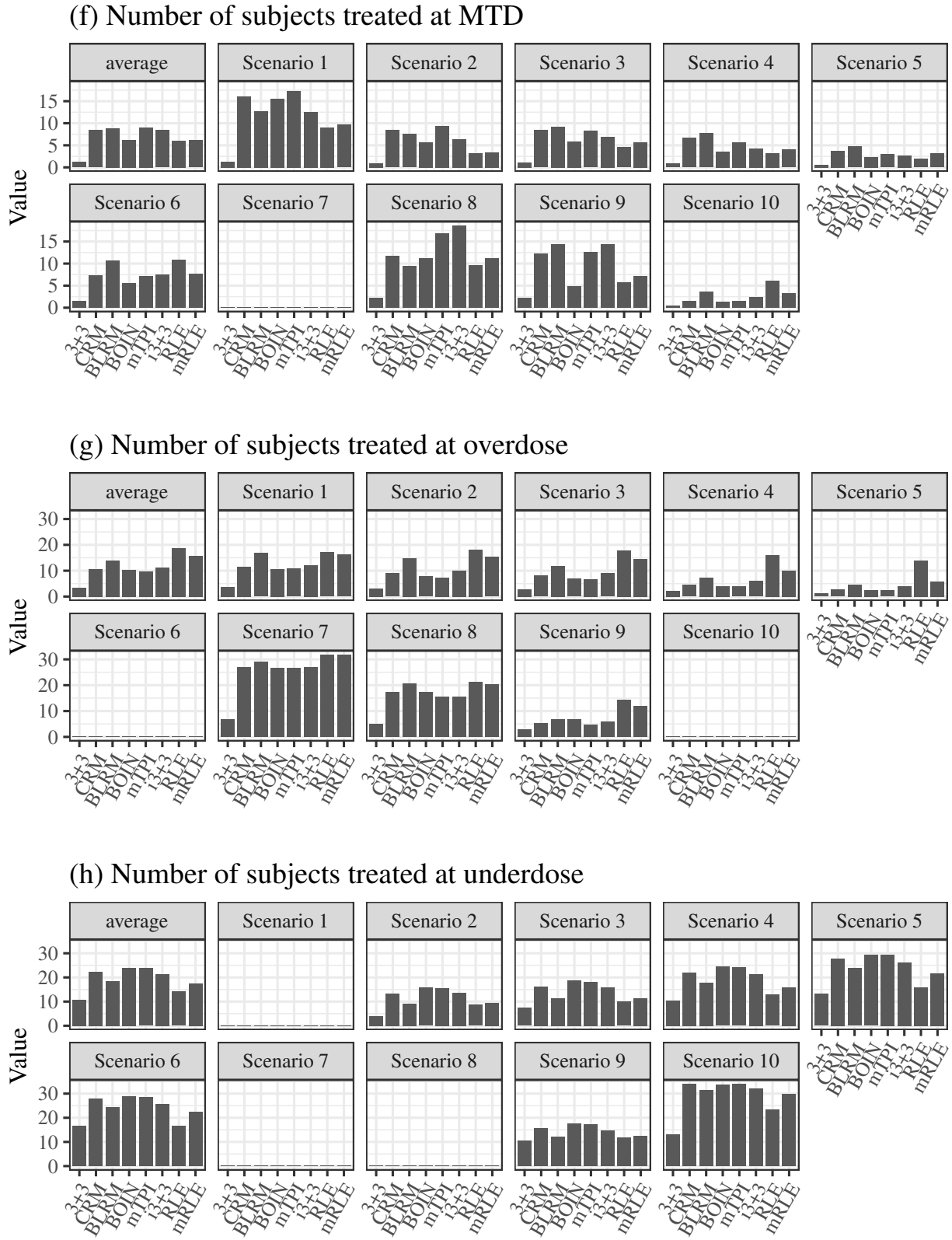

Figure 3.5: The results for performance metrics (f)–(h).

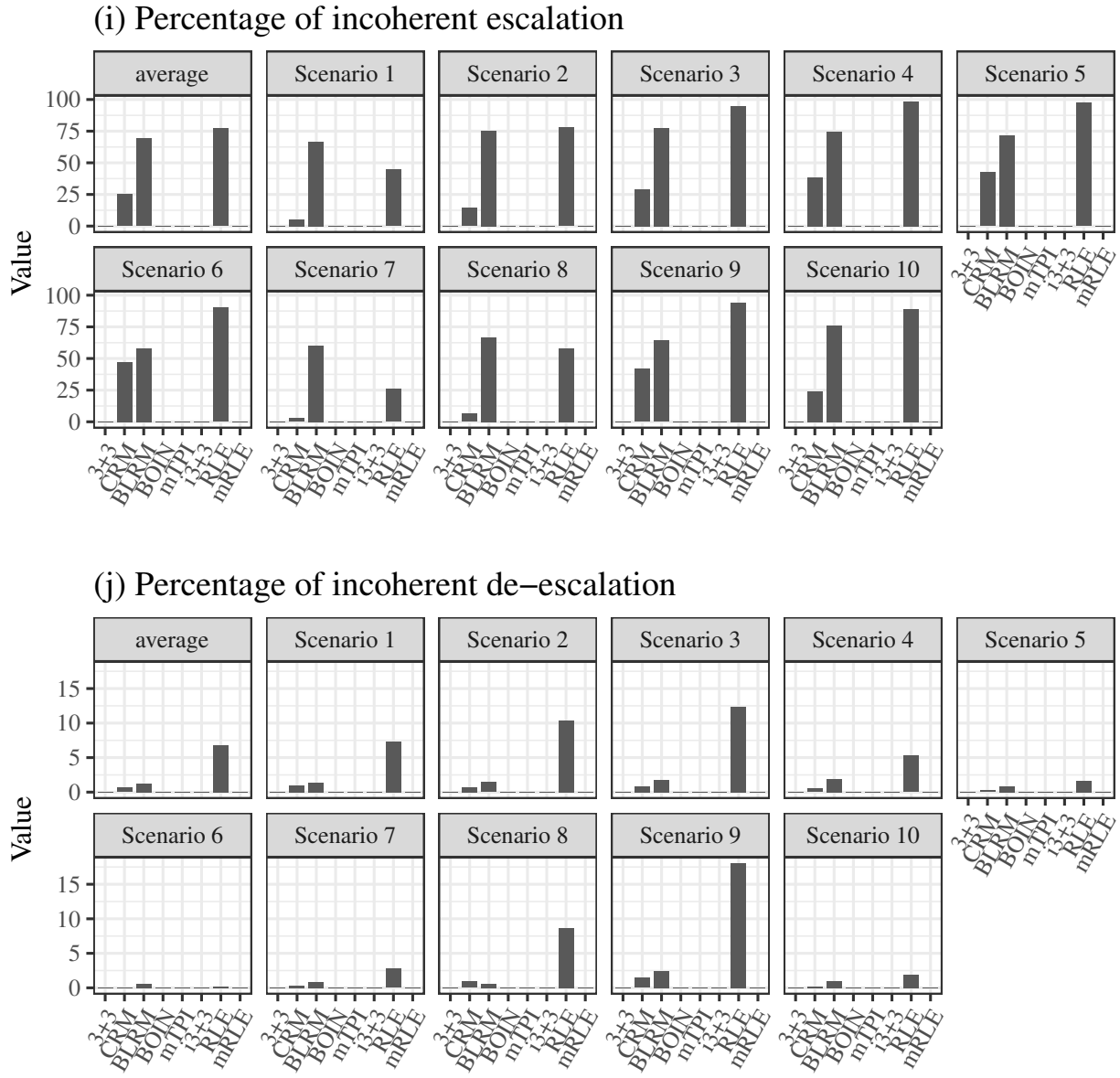

Figure 3.6: The results for performance metrics (i) and (j).

Table 3.3: Two examples of dose escalation paths in a single trial.

Case	Cohort	Dose Level	No. of Pats	DLTs	Decision
	1	1	3	2	stay
	2	1	3	0	stay
1	3	1	3	0	escalate
	4	2	3	0	escalate
	5	3	3	1	escalate
	1	1	3	0	escalate
	2	2	3	2	de-escalate
2	3	1	3	0	escalate
	4	2	3	0	stay
	5	2	3	0	escalate

In Cohort 5 of Case 1, *escalate* was selected in a situation where the DLTs were 2/9, 0/3, and 1/3 for Dose Level 1, 2, and 3, respectively. Since the DLT was 0/3 at Dose Level 2 and the probabilities of 0/3 and 1/3 are both 0.422 when $p_j = 0.25$ is true, RLE may have recommended Dose Level 4. This decision increased the percentage of incoherent escalation. However, such an aggressive decision might not be accepted by clinicians in practice. mRLE prohibited such a decision with ad-hoc rules, as explained in Section 3.3.1.

In Case 2, we compared mRLE with BOIN and i3+3. In Cohort 4 of Case 2, mRLE and i3+3 selected *stay* while BOIN selected *de-escalate*. Note that *stay* was not always selected when the DLT was 2/6 because the proposed methods consider the results at all dose levels. In Cohort 5 of Case 2, mRLE selected *escalate* while i3+3 selected *stay*. Thus, mRLE still shows more aggressive dose escalations than the other methods.

3.4 Summary

We have shown that deep reinforcement learning with an appropriately defined state, action, and reward can be used to construct an action selection rule that improves the PCS. In general, reinforcement learning becomes difficult when the reward (i.e., the correctness of the dose determined as the MTD evaluated at the end of each trial) is delayed, and the observation is noisy. Phase I oncology trials usually have these difficulties, and it is not obvious whether reinforcement learning works well in such situations. Nonetheless, we have shown that it can work successfully if we design and choose appropriately the Markov decision process as well as the learning algorithm and hyperparameters. Although this study describes a design with a single drug, the proposed method may be applied to clinical trials with multiple drugs by extending the state and action.

Chapter 4

Discussion

4.1 Optimal Adaptive Allocation in Dose-Response Studies for Non-Anticancer Drugs

One of the reasons why the existing methods did not work is that they are based on asymptotic theory, even though the actual sample size is of the order of 100. A method for a similar problem setting is the multi-armed bandits theory. However, it cannot be applied directly because (1) it assumes that all arms are mutually independent, (2) it assumes that the sample size is large enough, and (3) it optimizes a simple metric such as the sum of responses of all subjects rather than complex metrics such as power, TD, or MAE. We tried an approach that extended the multi-armed bandits theory by combining Monte Carlo tree search, but it did not improve the performance metrics. Reinforcement learning was found to be able to improve the metrics even with small sample sizes by making effective use of the information available. We consider that the function $Q_\pi(s, k)$, which represents the value of state s, is smooth with respect to s, so the smooth structure can be captured by the DNN.

A limitation of our method is that it is difficult to visualize and understand the obtained allocation rule intuitively because the state is multidimensional. An allocation example of RL-MAE in a single simulated trial is shown in Figure A.8 in Appendix A. Interactive software such as a Shiny application may help team members to understand the rule. In a clinical trial protocol, it is necessary to specify the assumptions (state, action, and reward) and the selected performance metric, and it would be helpful to show allocation examples.

In the definition of state, we used the differences from the placebo (e.g., $\bar{Y}_2 - \bar{Y}_1$) to avoid making assumptions about the placebo response. When we have a specific prior distribution reflecting the background knowledge on the placebo response, it is also natural to define the state by

$$s = \left\{ \bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_K, \hat{\sigma}_1, \ldots, \hat{\sigma}_K, \frac{n_1}{N}, \ldots, \frac{n_K}{N} \right\}.$$

We also simulated the proposed methods with slightly modified states, rewards, and model probabilities, which retrieved in general similar results. Nonetheless, it may be possible to slightly improve the performance by further tuning these settings.

Although the simulation study was conducted assuming Gaussian noise with the MCP-Mod method, the proposed method can also be applied to other settings (e.g., binary response) and other methods

(e.g., ANOVA and BMA). For example, if the variance of the observation noise is unknown, we will assume the prior distribution of the variance to generate it in reinforcement learning. Because the settings are quite standard in practice, we can expect that the proposed method can cover a wide range of actual clinical trials.

Results showed that the proposed methods was required to adjust the significance level to control the type I error rate. Therefore, developing a statistical test that is theoretically guaranteed under adaptive allocation is an important research topic. The optimization of power (RL-power) did not necessarily lead to improvements in other performance metrics. On the other hand, RL-MAE showed good results not only for MAE, but also for other metrics. This seems to intuitively correspond to the fact that if the dose-response curve itself is estimated with a small error, then the other purposes are achievable. For this reason, it seems natural to use RL-MAE if the focus is not on any particular metric. Note that it is theoretically the best to allocate all subjects to 0 and 8 mg to maximize the power under the scenarios used in the learning. RL-power indeed allocated many subjects to 0 and 8 mg, but there still exists a gap from this ideal allocation, which may be due to incomplete learning. Therefore, further tuning of the parameters and neural network may improve the performance. The results described in Appendix A showed that RL-MS and RL-MAE performed poorly when the exponential model was true. This indicates that if the candidate models considerably differ from the true model, the allocation rules obtained from the learning may not perform well. It may be important to specify the distribution that will generate many possible models in reinforcement learning. In fact, by including the exponential model in the learning, we obtained good performance without sacrificing the performance for the other models (see Appendix A).

When the proposed method is used, the number of subjects allocated could be unbalanced. When the imbalance must be taken into consideration for safety or ethical reasons, the number of subjects allocated equally at the beginning of the trial can be increased or a penalty can be incorporated in the reward if the number of subjects at a dose does not reach the threshold. Although we allocated subjects probabilistically according to the discrete distribution $\pi^*(k|s)$ when applying the obtained rule, we can also use the rounding method, such as the one reported by Pukelsheim and Rieder (1992). The results were generally similar.

Although we used one performance metric for optimization, any metric can be used, including a combination of multiple metrics, because the method does not depend on the specific properties of the performance metric. Since many factors other than dose-response are involved in actual phase III trials, it is also important to develop an appropriate performance metric for the success of phase III trials. Furthermore, we believe that it is possible to extend this approach for the stopping rules for success or futility by adding a stopping option as one of the actions and defining an appropriate reward for the option. It remains to be verified in which situations this will apply.

4.2 Optimal Dose Escalation Methods in Dose-Finding Studies for Anticancer drugs

One of the reasons why the PCS of the existing methods are not high is that their dose-escalation rules are very simple and do not aim to maximize the PCS. We consider that the proposed method enables efficient MTD search through reinforcement learning.

As shown in Section 3.3.4, the dose escalation path using the obtained rule may be interpretable. A limitation of the proposed method is that it is not always possible to interpret why the obtained

rule selected a decision in a given situation. In practice, we believe that the dose in the next cohort should be determined through discussion with the clinical team, taking into account the decision recommended by the proposed method.

The simulation study was conducted in a setting that was expected for a common drug. Because the settings used are quite standard in practice, we can expect that the proposed method can cover a wide range of actual clinical trials. We also evaluated the methods in extreme scenarios such as $(0.1, \ldots, 0.1)$ and $(0.9, \ldots, 0.9)$; the former result was similar to Scenario 6, and the latter result was similar to Scenario 7. The results in Section 3.3 show that RLE improved the PCS compared with other methods. On the other hand, RLE had fewer subjects assigned to the MTD and a higher number of DLTs. However, since reducing the number of DLTs may worsen performance metrics (a)–(c), it is necessary to consider the clinical significance of each metric and which metric should be prioritized. Regardless of how we adjusted the settings, the other methods could not increase PCS at the expense of DLT. Therefore, we believe that our method offers a good option in situations where increasing PCS is important. It is also possible to make escalation in the learned rule less aggressive by setting negative rewards for the number of DLTs and incoherent escalations. This will be the focus of future work.

We also simulated the proposed method with slightly modified state and action options (e.g., removing the last element $\sum_{j=1}^{J} x_j/N$ of state s, or adding an action to determine no MTD and stop the trial), and the results were generally similar. Because the results of reinforcement learning depend on the doses and scenarios, further research and simulations are needed to evaluate the superiority of RLE.

To facilitate reinforcement learning, the proposed method set the scenarios in which the MTD was clearly defined from J, ϕ, ε, and δ. However, there is room for improvement in how we generate scenarios for learning. When we added small Gaussian noise to the 20 scenarios in Table 3.2, the results were generally similar. When we generated scenarios using the pseudo-uniform algorithm of Clertant and O'Quigley (2017) and conducted reinforcement learning, the PCS could not be improved for realistic scenarios, including the 10 scenarios in Table 3.1. This may be because the distribution of scenarios generated by the algorithm deviates from the realistic scenario distribution based on background knowledge. We believe that it is important to evaluate various scenarios before the start of the trial.

Chapter 5

Conclusion

This study examined optimal adaptive allocation using deep reinforcement learning in dose-finding studies. In dose-finding studies for both non-anticancer and anticancer drugs, the issue was that the allocation of subjects to each dose has not been determined based on the performance metrics to be optimized.

For the issue, we showed that deep reinforcement learning with an appropriately defined state and reward can be used to construct adaptive allocation rules that can directly optimize the performance metric to be optimized. In general, reinforcement learning becomes difficult when the reward (i.e., the performance metric evaluated at the end of each trial) is delayed, and the observation is noisy. Clinical trials have these difficulties, and it is not obvious whether reinforcement learning works successfully to address the same. Nonetheless, we have shown that it can work well if we appropriately design and choose the Markov decision process as well as the learning algorithm and hyperparameters.

In dose-finding studies for non-anticancer drugs, we used deep reinforcement learning to construct an adaptive subject allocation rule to directly optimize each performance metric (Matsuura et al., 2022). Specifically, by using deep reinforcement learning with an appropriately defined state, reward and environment, we numerically derived an adaptive allocation rule, which is a discrete distribution $\pi^*(k|s)$ of taking a dose k under state s). In the proposed method, state s was defined as a vector of the mean and standard deviation of the responses of the subjects allocated to each dose, and the number of subjects allocated to each dose. A reward was defined using one of the above metrics. For the environment, we used the assumptions of the model in the method to estimate the dose-response curve. We conducted a simulation study in a slightly modified setting used by Bornkamp et al. (2007) and Dragalin et al. (2010). We compared the performance of the equal allocation, D-optimal method, TD-optimal method, and proposed method. To access the effect of different allocation, the method to estimate the dose-response curve was fixed to the MCP-Mod method. It was shown that the proposed method improved the performance metric to be optimized, the other performance metrics also improved when the performance metric to be optimized was the mean absolute error, and the performance was kept even in scenarios that deviated from the environment used in reinforcement learning.

In dose-finding studies for anticancer drugs, we used deep reinforcement learning to construct an action selection rule to directly optimize the PCS (Matsuura et al., 2023). Specifically, by using deep reinforcement learning with an appropriately defined state, action, reward and environment, we numerically derived an action selection rule, which was a discrete distribution $\pi^*(a|s)$ of taking

an action a under state s). In the proposed method, state s was defined as a vector of the previous dose level, the proportion of the number of subjects assigned, and the proportion of the number of DLTs. Action a was defined as a vector of "continue the trial and escalate/stay/de-escalate the dose in the next cohort" or "discontinue the trial". A reward was defined so that they could be obtained only when the MTD was correctly estimated. For the environment, we used hypothetical scenarios in which the MTD was clearly defined to stably learn π^*. We conducted a simulation study in a slightly modified setting used in Appendix C of Zhou et al. (2018b). We compared the performance of the 3+3 design, CRM, BLRM (Neuenschwander et al., 2008), BOIN, mTPI (Ji et al., 2010), i3+3 (Liu et al., 2020), and proposed methods. The proposed method was shown to improve the PCS over the existing methods, although it caused slightly more toxicity during the trial. In particular, the proposed method stably performed better in the scenario where the MTD was on the high-dose level, as is likely to be more frequent in actual clinical trials.

We believe that these results will contribute to improving the probability of drug approval through the improvement of performance metrics in dose-setting studies.

Appendix A

This appendix shows figures that supplement the results of the simulation study in Chapter 2 (e.g., results for each model and an allocation example of the proposed method).

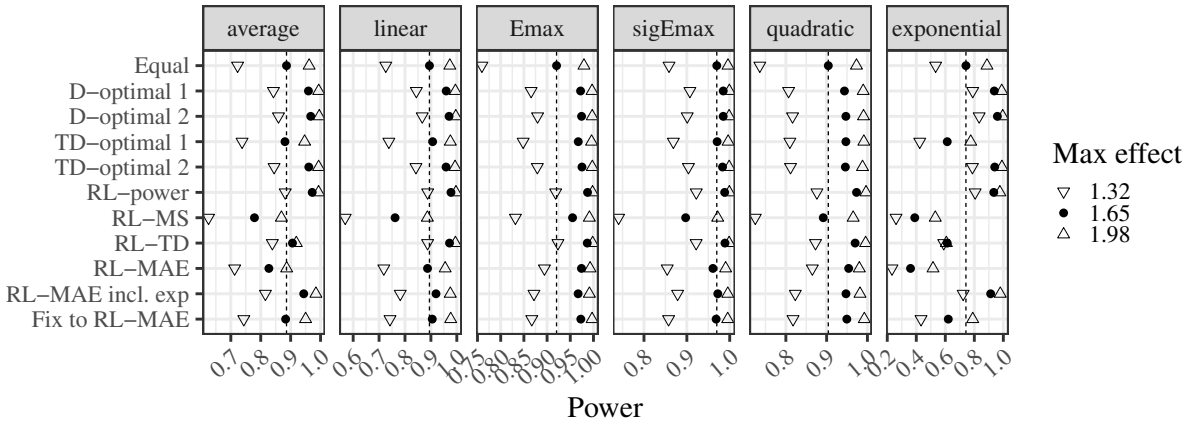

Figure A.1: The results for power. RL-MAE incl. exp represents that simulated data in reinforcement learning is generated from the four models (linear, Emax, sigEmax, and exponential) with equal probabilities. Fix to RL-MAE represents a fixed design with the number of subjects equal to the average of those of RL-MAE.

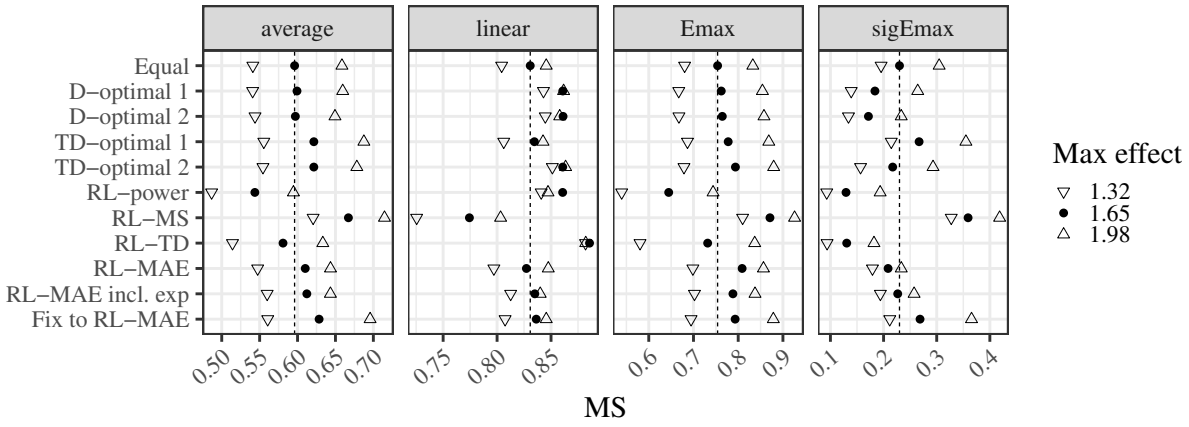

Figure A.2: Probability of selecting the true model.

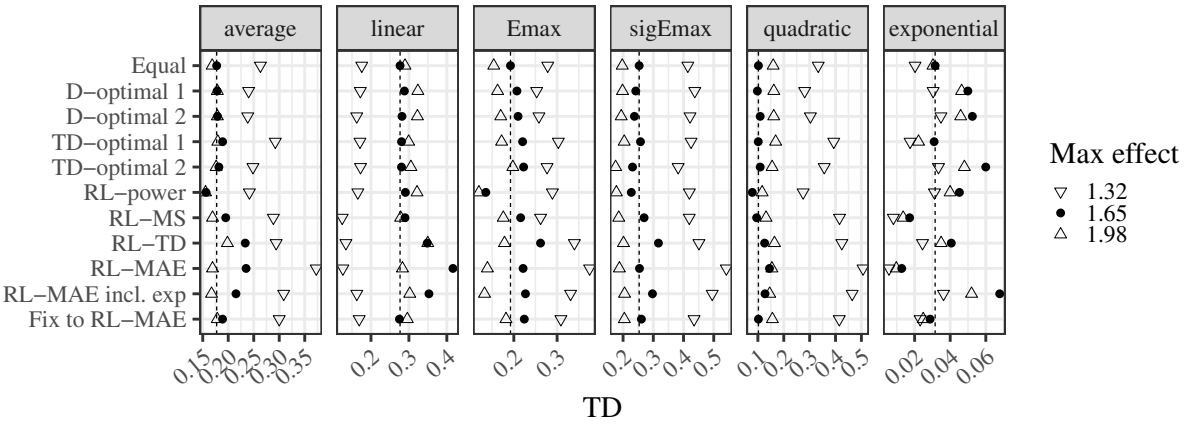

Figure A.3: Probability that the estimated target dose is within the interval $I^d_{\text{targ}}(0.1)$.

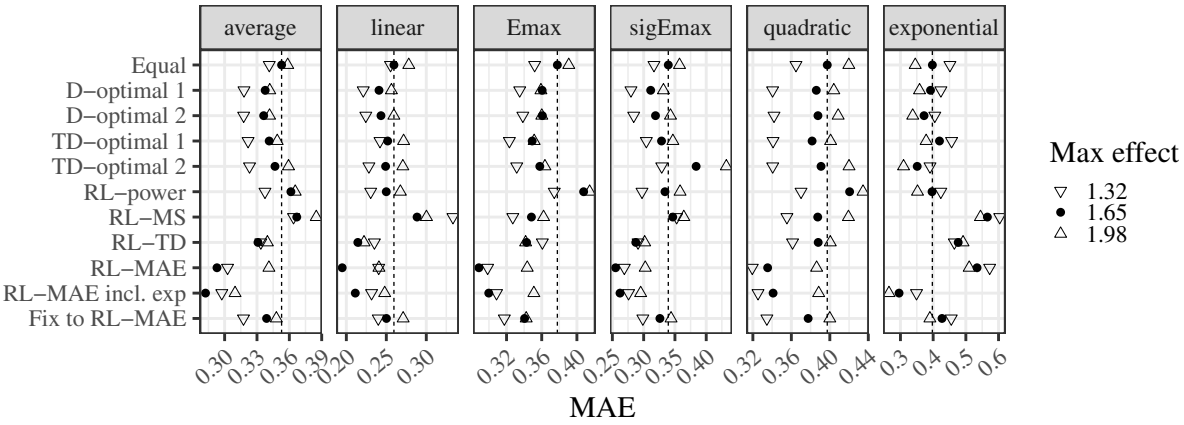

Figure A.4: The results for MAE. Smaller MAE implies better accuracy.

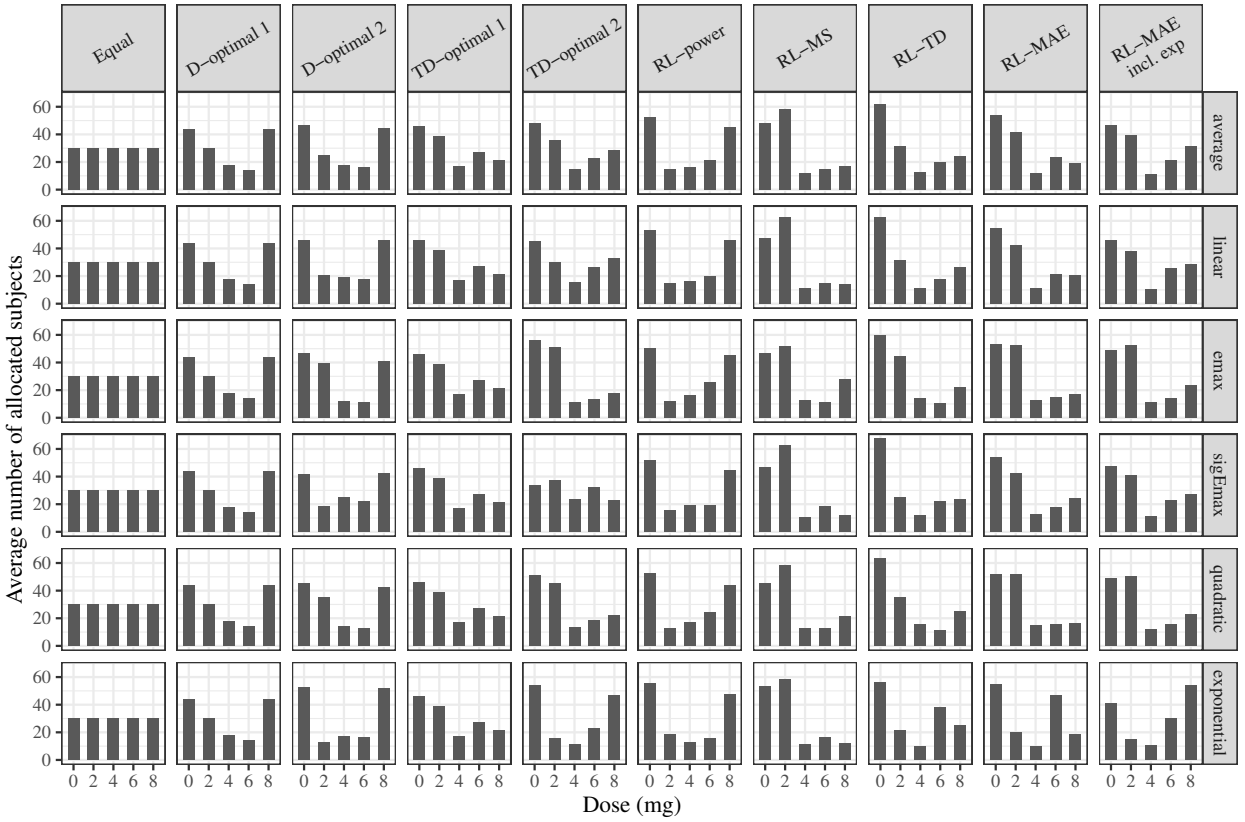

Figure A.5: The results for the average number of subjects allocated when the maximum effect was 1.65.

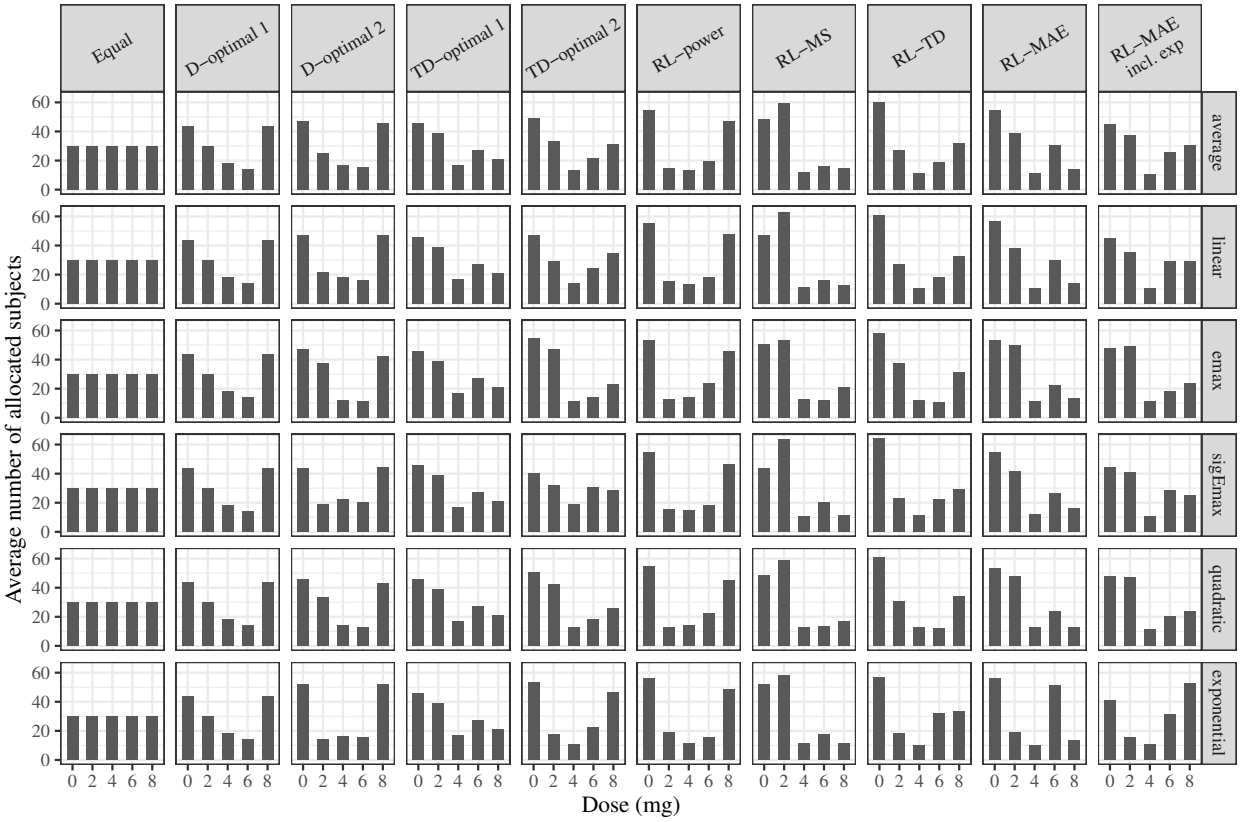

Figure A.6: The results for the average number of subjects allocated when the maximum effect was 1.32.

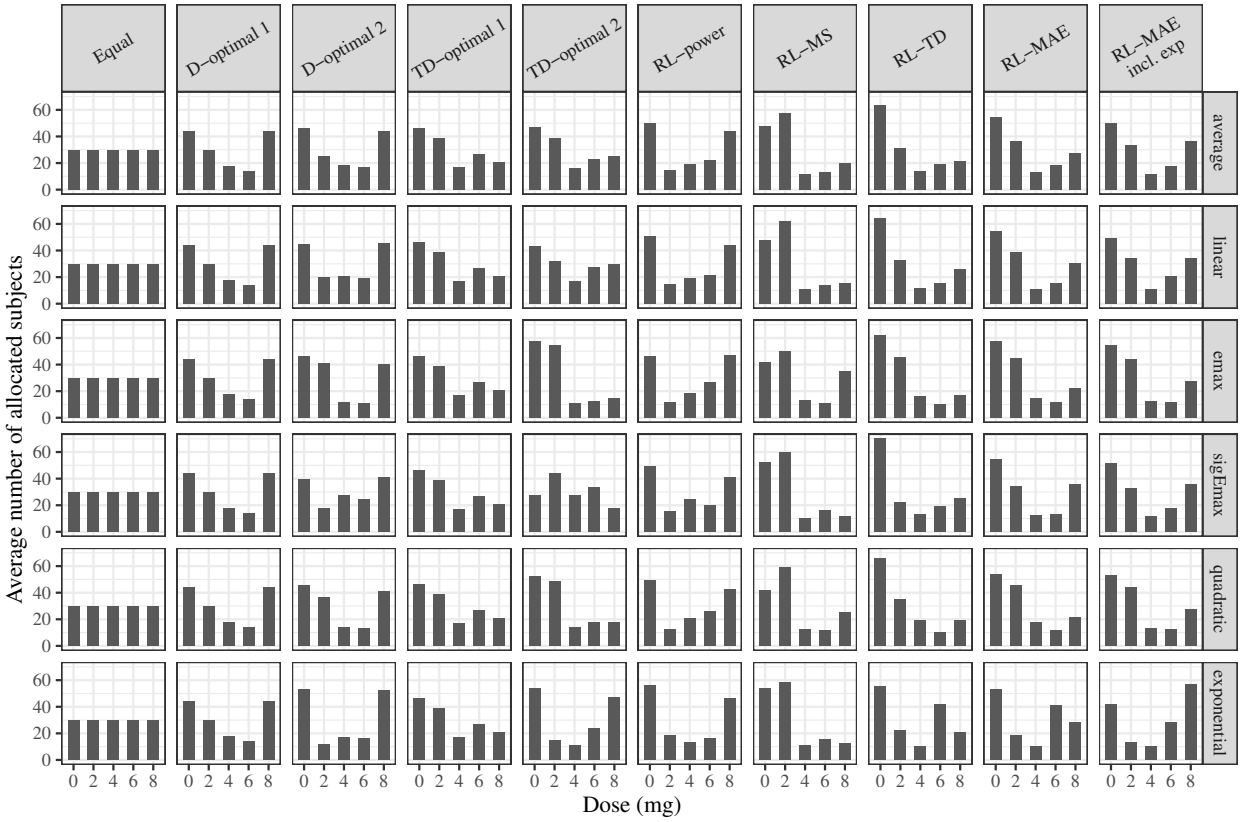

Figure A.7: The results for the average number of subjects allocated when the maximum effect was 1.98.

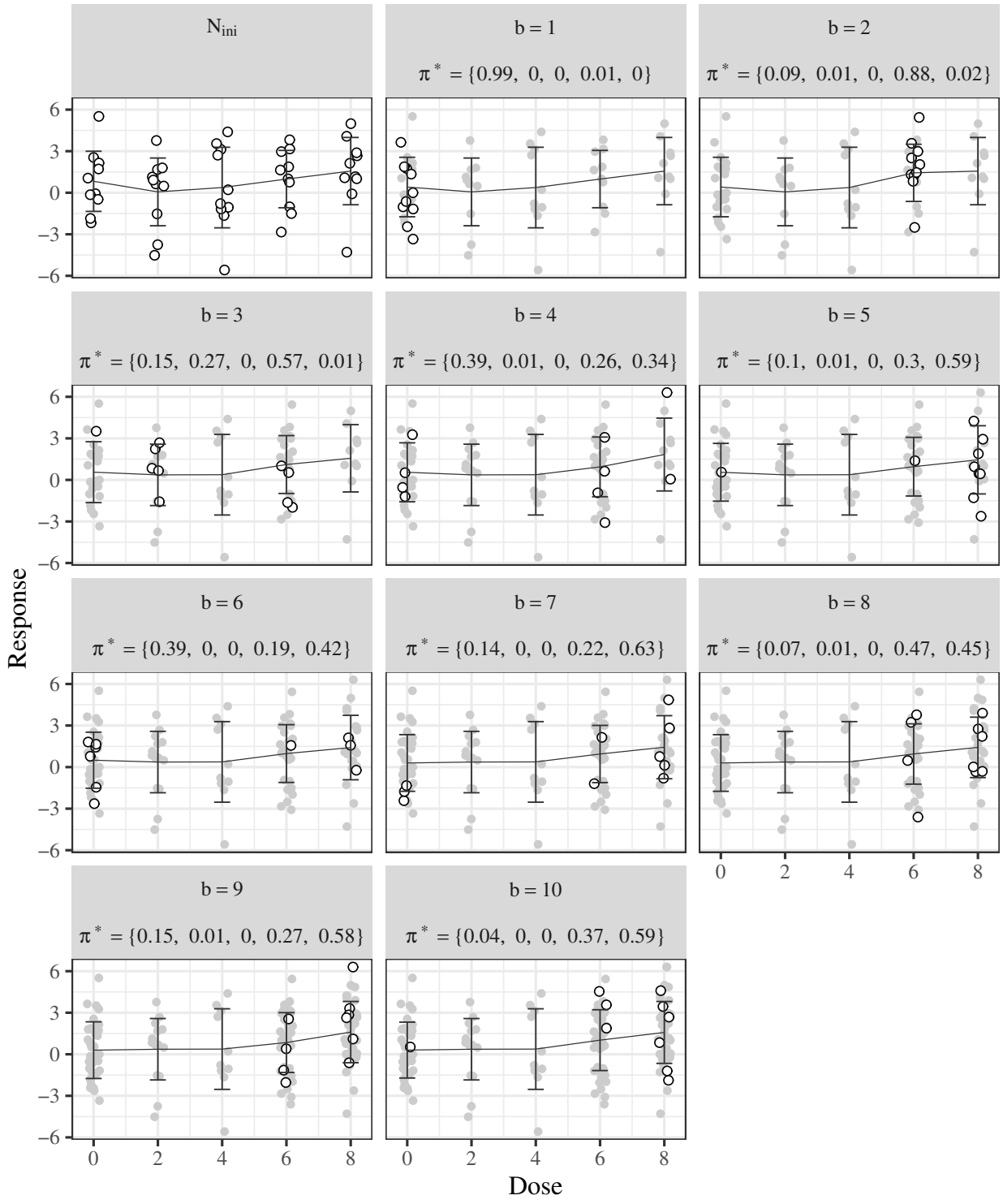

Figure A.8: An allocation example of RL-MAE in a simulated trial when the linear model was true and the maximum effect was 1.65. b represents the situation after allocating the subjects in the b-th block and obtaining their responses. The values of $\pi^*(k|s)$ were calculated before allocating the subjects in the b-th block. The open circles represent the responses of the subjects allocated in the b-th block. The gray circles represent the responses of the subjects allocated in the previous blocks. The black lines represent the means of the responses at each dose, and the error bars represent ± 1 SD. At the end of this simulated trial, the MCP-Mod method calculated the p-value as 0.0067 and selected the linear model.

Appendix B

In this appendix, we introduce the overview of reinforcement learning techniques used in the proposed method. See, e.g., Sutton and Barto (2018) for general introduction to reinforcement learning.

Preliminaries for Reinforcement Learning

Formally, a reinforcement learning problem can be characterized by a Markov Decision Process defined by the 4-tuple $(\mathcal{S}, \mathcal{A}, P, R)$:

- State space of the environment $\mathcal{S}$: At each time step t the agent observes a state of the environment denoted $s_t \in \mathcal{S}$. The initial state is sampled from an initial distribution on $\mathcal{S}$.
- Action space $\mathcal{A}$: At time step t, the agent selects an action $a_t \in \mathcal{A}$ according to a *policy* π as $a_t \sim \pi(\cdot|s_t)$, where $\pi(\cdot|s)$ is a probability distribution over $\mathcal{A}$ that represents the strategy of the agent when the state is s.
- Transition probability $P(s'|s, a)$: Given the action a_t and the state s_t, the environment evolves into a new state s_{t+1} with probability $P(s_{t+1}|s_t, a_t)$.
- Reward function $R(s, a, s')$: The agent receives a reward $r_t = R(s_t, a_t, s_{t+1})$ when taking action a_t at state s_t and the new state becomes s_{t+1}. We denote by $r(s_t, a_t)$ the expected value of $R(s_t, a_t, s_{t+1})$ given (s_t, a_t).

The interaction between the agent and the environment lasts for an episode, that is limited by time or by reaching a terminal state, and then the process restarts. For simplicity of notation we denote by T the end of the interaction and it can be finite or infinite.

The return G_t defined by

$$G_t = \sum_{k=t}^{T} \gamma^{k-t} r(s_k, a_k)$$

is the discounted cumulative reward after time t with a discount factor $\gamma \in [0, 1]$. Small values of γ leads the agent to focus on short-term rewards while a large value favors long-term rewards. The goal of the agent is to select at each state the action that will lead the highest expected cumulative discounted reward.

The value function V_π of a state s is the expected return from this state following the policy π and denoted as

$$V_\pi(s) = \mathbb{E}_\pi(G_t|s_t = s),$$

where the expectation is taken over all possible stochastic trajectories under π. The state-action value function Q_π is the expected return from state s by taking action a and after that following the policy π:

$$Q_\pi(s,a) = \mathbb{E}_\pi(G_t|s_t = s, a_t = a).$$

The advantage function A_π is defined by

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s).$$

The advantage expresses how much better or worse the reward obtained by action a in state s is compared with the average expected reward $V_\pi(s)$ from state s.

Proximal Policy Optimization

Many algorithms have been proposed to optimize π. For example, deep Q-network (DQN) tries to find a good policy indirectly by estimating Q-values and choosing the action maximizing the estimated Q-value for each state s (Mnih et al., 2015). In contrast, policy optimization methods perform a gradient update directly on the parameters of a policy.

Proximal Policy Optimization (PPO) belongs to these methods, which parameterizes the policy as $\pi(a_t|s_t;\theta)$ and update θ directly using the observed rewards r_t. For update of θ, PPO considers the following objective function to optimize in order to estimate the optimal parameter $\hat{\theta}$.

Let us denote the probability ratio between the old and new policies as

$$f(\theta,t) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)},$$

where θ_{old} is the parameter obtained in the last update. Conceptually, PPO tries to improve the policy by maximizing

$$J(\theta) = \mathbb{E}_t(f(\theta,t)\hat{A}_{\theta_{\text{old}}}(s_t,a_t)), \tag{B.1}$$

where $\mathbb{E}_t$ indicates the empirical average over a finite batch of trajectories (the histories of the states, actions, and rewards through timesteps) and $\hat{A}_{\theta_{\text{old}}}(s_t,a_t)$ is an approximation of the advantage function (this approximation is discussed later).

Still, it is empirically known that maximizing this objective with respect to θ, without a restriction on the distance between θ_{old} and θ, results in instability and too aggressive updates. PPO solves this problem by imposing a constraint on $f(\theta,t)$ to be within a small interval around 1, precisely $[1-\epsilon, 1+\epsilon]$. To be more specific, PPO replaces $f(\theta,t)\hat{A}_{\theta_{\text{old}}}(s_t,a_t)$ in Eq. (B.1) with

$$\min\{f(\theta,t)\hat{A}_{\theta_{\text{old}}}(s_t,a_t), \text{clip}(f(\theta,t), 1-\epsilon, 1+\epsilon)\hat{A}_{\theta_{\text{old}}}(s_t,a_t)\},$$

where $\text{clip}(x,a,b) = \min\{\max\{x,a\},b\}$ for $a < b$.

The estimator of $\hat{A}_{\theta_{\text{old}}}(s_t,a_t)$ is built using generalized advantage estimation (GAE) (Schulman et al., 2015) given by

$$\hat{A}_{\theta_{\text{old}}}(s_t,a_t) = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1},$$

where $\lambda \in [0, 1]$ is a hyperparameter and $\delta_t = r_t + \gamma V_{\theta_{\text{old}}}(s_{t+1}) - V_{\theta_{\text{old}}}(s_t)$.

Here note that the estimation of the advantage function involves the learned value function $V_\theta(s)$. The authors of PPO suggest to use the Actor-Critic method to estimate the value function. As the name suggest, it has two components: the actor and the critic. The actor corresponds to the policy π and is used to choose the action for the agent. The critic corresponds to the value function V. The Actor-Critic is represented by a shared neural network θ, which then branches into two heads (one for the actor and one for the critic) at the end of the architecture (Figure B.1). Therefore θ includes the parameter θ_v for the value function V and θ_π for the policy π. So that the critic approximates the actual return well, PPO imposes the penalty for the approximation error given by

$$-(V_\theta(s_t) - V_t^{\text{target}})^2,$$

where $V_t^{\text{target}} = G_t$ is the observed return (from the simulation) at time t, $V_\theta(s_t)$ is the estimated value function from the neural network at state s_t.

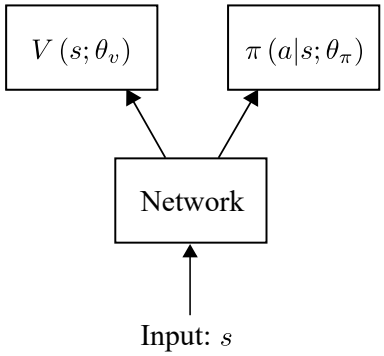

Figure B.1: Actor-Critic PPO network

In summary, PPO maximizes the following objective function:

$$\begin{aligned} J(\theta)^{\text{CLIP}} = \mathbb{E}_t\big(\, &\min\{f(\theta,t)\hat{A}_{\theta_{\text{old}}}(s_t,a_t), \text{clip}(f(\theta,t), 1-\epsilon, 1+\epsilon)\hat{A}_{\theta_{\text{old}}}(s_t,a_t)\} \\ &- c_1(V_\theta(s_t) - V_t^{\text{target}})^2 + c_2 S[\pi_\theta](s_t)\big), \end{aligned}$$

where c_1, c_2 and ϵ are hyperparameters. Here S in the last term is some entropy function such as $S[\pi](s) = -\sum_{a\in\mathcal{A}} \pi(a|s)\log\pi(a|s)$, which gives a bonus to a policy that explores a variety of actions.

Algorithm

For each update of the neural network parameter θ, an actor collects data of T_{train} timesteps over multiple episodes. Then we construct the objective function given in the last section on these T_{train} timesteps data, and maximize it with the stochastic gradient descent (SGD) or Adam algorithm.

Specification in Our Problems

In our problem setting in Chapters 2 and 3, the episode, time t, and policy π correspond to the simulated trial, block b, and allocation rule π, respectively. For the reward, r had a value described

Algorithm 1 PPO Algorithm

procedure PPO($\mathcal{S}, \mathcal{A}, P, R$)
  Initialize weights θ_{old}.
  while the total number of episodes $\leq E$ **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T_{train} timesteps, where a new episode starts from the initial state when the current episode has ended.
    Collect a batch of T_{train} samples $\{(s_i, a_i, r_i, s_i')\}_{i=1}^{T_{\text{train}}}$.
    Compute advantage $\hat{A}_{\theta_{\text{old}}}(s_t, a_t)$ for T_{train} timesteps.
    Update the neural network parameters $\theta_{\text{old}} \leftarrow \underset{\theta}{\text{argmax}}\, J(\theta)^{\text{CLIP}}$.
  end while
end procedure

in Sections 2.2.3 and 3.2.2 if the time was at the end of the trial, but otherwise $r = 0$. We used the discount factor $\gamma = 1$ (that is, there is no discount). Therefore, the expected return is equivalent to the expected value of the reward at the end of the trial. For the advantage, we used $\lambda = 1$. For the objective function, we used $c_1 = 1$, $c_2 = 0$, and $\epsilon = 0.3$. For the update of θ, we used $E = 1{,}000{,}000$ and $T_{\text{train}} = 10{,}000$ in Chapter 2. That is, θ was updated after each $T_{\text{train}}/B = 10{,}000/10 = 1{,}000$ simulated trials. We used $E = 3{,}000{,}000$ and $T_{\text{train}} = 10{,}000$ in Chapter 3. To maximize the object function, we used the SGD algorithm with the minibatch size = 200, the stepsize (i.e., learning rate) = 0.00005, and the number of epochs = 20. Although we used the default values of the software for λ, c_1, c_2, ϵ, and the stepsize, we had to tune T_{train} and the minibatch size.

Acknowledgments

I am deeply grateful to Professor Takashi Sozu for teaching me how to be a researcher in biostatistics, and giving several insightful comments on this thesis.

I have greatly benefited from Dr. Junya Honda in Kyoto University, Dr. Kentaro Sakamaki in Yokohama City University, and Mr. Imad El Hanafi for elaborated guidance and thorough review of their co-authored manuscripts.

I would like to thank Dr. Shuji Ando, Dr. Jun Tsuchida, Dr. Tomohiro Shinozaki, and Mr. Koichi Hashizume for their useful comments in the discussions.

I also express my appreciation to Professors Takako Akakura, Tohru Ikeguchi, Go Irie, and Hiroki Hashiguchi of Tokyo University of Science. All were members of my doctoral committee and provided many helpful comments and suggestions.

My final acknowledgement is to my family (including my three cats) who warmly encourage and relax me.

References

1. Aouni, J., Bacro, J.N., Toulemonde, G., et al. (2020) Design optimization for dose-finding trials: A review. *Journal of Biopharmaceutical Statistics*, 30 (4): 662–673.

2. Bornkamp, B., Bretz, F., Dette, H., et al. (2011) Response-adaptive dose-finding under model uncertainty. *The Annals of Applied Statistics*, pp. 1611–1631.

3. Bornkamp, B., Bretz, F., Dmitrienko, A., et al. (2007) Innovative approaches for designing and analyzing adaptive dose-ranging trials. *Journal of Biopharmaceutical Statistics*, 17 (6): 965–995.

4. Bornkamp, B., Pinheiro, J. and Bretz, F. (2009) MCPMod: An R package for the design and analysis of dose-finding studies. *Journal of Statistical Software*, 29 (7): 1–23.

5. Bretz, F., Hsu, J., Pinheiro, J., et al. (2008) Dose finding–a challenge in statistics. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50 (4): 480–504.

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

7. Carpenter, B., Gelman, A., Hoffman, M.D., et al. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software*, 76 (1).

8. Clertant, M. and O'Quigley, J. (2017) Semiparametric dose finding methods. *Journal of the Royal Statistical Society: Series B*, 79 (5): 1487–1508.

9. Daimon, T., Hirakawa, A. and Matsui, S. (2019) *Dose-finding designs for early-phase cancer clinical trials: A brief guidebook to theory and practice*. New York: Springer.

10. Dette, H., Bretz, F., Pepelyshev, A., et al. (2008) Optimal designs for dose-finding studies. *Journal of the American Statistical Association*, 103 (483): 1225–1237.

11. Dragalin, V., Bornkamp, B., Bretz, F., et al. (2010) A simulation study to compare new adaptive dose-ranging designs. *Statistics in Biopharmaceutical Research*, 2 (4): 487–512.

12. Dragalin, V., Hsuan, F. and Padmanabhan, S.K. (2007) Adaptive designs for dose-finding studies based on sigmoid emax model. *Journal of Biopharmaceutical Statistics*, 17 (6): 1051–1070.

13. EMA (2014) *Qualification opinion of MCP-mod as an efficient statistical methodology for model-based design and analysis of phase II dose finding studies under model uncertainty*.

14. Espeholt, L., Soyer, H., Munos, R., et al. (2018) IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. *Proceedings of the 35th International Conference on Machine Learning*, 80: 1407–1416.

15. FDA (2015) *FDA qualification of MCP-MOD method*.

16. FDA (2017) *PDUFA reauthorization performance goals and procedures fiscal years 2018 through 2022*.

17. Fujimoto, S., Hoof, H. and Meger, D. (2018) Addressing function approximation error in actor-critic methods. *Proceedings of the 35th International Conference on Machine Learning*, 80: 1587–1596.

18. Gould, A.L. (2019) BMA-mod: A bayesian model averaging strategy for determining dose-response relationships in the presence of model uncertainty. *Biometrical Journal*, 61 (5): 1141–1159.

19. Green, S., Benedetti, J., Smith, A., et al. (2012) *Clinical trials in oncology, 3rd edition*. New York: Chapman; Hall/CRC.

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

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

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

23. Lewis, R.J. and Berry, D.A. (1994) Group sequential clinical trials: A classical evaluation of bayesian decision-theoretic designs. *Journal of the American Statistical Association*, 89 (428): 1528–1534.

24. Liang, E., Liaw, R., Nishihara, R., et al. (2018) RLlib: Abstractions for distributed reinforcement learning. *Proceedings of the 35th International Conference on Machine Learning*, 80: 3053–3062.

25. Liu, M., Wang, S.-J. and Ji, Y. (2020) The i3+3 design for phase I clinical trials. *Journal of Biopharmaceutical Statistics*, 30 (2): 294–304.

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

27. Matsuura, K. (2022) *Optimal Adaptive Allocation using Deep Reinforcement Learning in a Dose-Response Study*. Available at: https://github.com/MatsuuraKentaro/Optimal_Adaptive_Allocation_in_a_Dose-Response_Study.

28. Matsuura, K. (2023) *Optimal Dose Escalation in Phase I Oncology Trials*. Available at: https://github.com/MatsuuraKentaro/Optimal_Dose_Escalation_in_Phase_I_Oncology_Trials.

29. Matsuura, K., Honda, J., El Hanafi, I., et al. (2022) Optimal adaptive allocation using deep reinforcement learning in a dose-response study. *Statistics in Medicine*, 41 (7): 1157–1171.

30. Matsuura, K., Sakamaki, Kentaro, Honda, J., et al. (2023) Optimal dose escalation methods using deep reinforcement learning in phase I oncology trials. *Journal of Biopharmaceutical Statistics*, (accepted).

31. Mercier, F., Bornkamp, B., Ohlssen, D., et al. (2015) Characterization of dose-response for count data using a generalized MCP-mod approach in an adaptive dose-ranging trial. *Pharmaceutical Statistics*, 14 (4): 359–367.

32. Miller, F., Guilbaud, O. and Dette, H. (2007) Optimal designs for estimating the interesting part of a dose-effect curve. *Journal of Biopharmaceutical Statistics*, 17 (6): 1097–1115.

33. Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015) Human-level control through deep reinforcement learning. *Nature*, 518 (7540): 529–533.

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

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

36. Ohlssen, D. and Racine, A. (2015) A flexible bayesian approach for modeling monotonic dose-response relationships in drug development trials. *Journal of Biopharmaceutical Statistics*, 25 (1): 137–156.

37. Pukelsheim, F. and Rieder, S. (1992) Efficient rounding of approximate designs. *Biometrika*, 79 (4): 763–770.

38. Schulman, J., Moritz, P., Levine, S., et al. (2015) High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*. Available at: http://arxiv.org/abs/1506.02438.

39. Schulman, J., Wolski, F., Dhariwal, P., et al. (2017) Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*. Available at: http://arxiv.org/abs/1707.06347.

40. Smietana, K., Siatkowski, M. and Møller, M. (2016) Trends in clinical success rates. *Nature Reviews Drug Discovery*, 15 (6): 379–80.

41. Sutton, R.S. and Barto, A.G. (2018) *Reinforcement learning: An introduction*. Cambridge: MIT press.

42. Wong, C.H., Siah, K.W. and Lo, A.W. (2019) Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20 (2): 273–286.

43. Yan, F., Mandrekar, S.J. and Yuan, Y. (2017) Keyboard: A novel bayesian toxicity probability interval design for phase I clinical trials. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 23 (15): 3994–4003.

44. Zhou, H., Murray, T.A., Pan, H., et al. (2018a) Comparative review of novel model-assisted designs for phase I clinical trials. *Statistics in Medicine*, 37 (14): 2208–2222.

45. Zhou, H., Yuan, Y. and Nie, L. (2018b) Accuracy, safety, and reliability of novel phase I trial designs. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 24 (18): 4357–4364.