Mutl-Armed Bandit and Response-Adaptive Randomisation

Rui-Yang Zhang

1 Introduction

The **multi-armed bandit problem** (MABP) considers the decision problem of optimally allocating resources (or pulls) to different arms in a sequential manner (making further decisions based on current observations rather than following the same rule throughout) to maximise the overall expected gain of pulling the arms (4). The MABP has been studied as a key example of a reinforcement learning (RL) problem (8) and the field of RL has received a significant amount of attention and development over the years due to successful applications such as AlphaGo (7). Applications of MABP in the design of **clinical trials**, although being one of the earliest applications in mind of MABP (e.g. Thompson Sampling (10)), have been rarely implemented in real life.

The gold standard of clinical trial design is the **randomisation controlled trial** (RCT) (3), where each patient is allocated to the control group or the treatment group with the same *fixed* and *pre-determined* probability. A consequence of this strict randomisation rule is that we might allocate patients to a particular group even when there is strong evidence that the other group has superior performance. This raises the ethical problem of individual benefit (giving the patient more effective treatment) versus collective benefit (maintaining the rigour of the trial) (2). Because of this issue, people have considered **response-adaptive randomisation** (RAR) (5) as an alternative which adjusts the randomisation probability according to the trial outcomes at various interim times.

However, the RAR trials have been considered *controversial* in the community. Some polar opposite opinions include: "If you are planning a randomized comparative clinical trial and someone proposes that you use outcome adaptive randomization, Just Say No" (9), and "... optimal [RAR] designs allow implementation of complex optimal allocations in multiple-objective clinical trials and provide valid tools to inference in the end of the trial. In many instances they prove superior over traditional balanced randomization designs in terms of both statistical efficiency and ethical criteria" (6).

A key failed attempt of RAR trials is the ECMO trial of (1). In that trial, due to the initial successes of the treatment, way too many patients were allocated to the treatment group, causing a significant imbalance in the number of patients in the two groups (11 out of 12 total patients were allocated to the treatment group). This raised serious doubts within the community regarding the reliability of the trial results, causing a second trial

to be conducted using fixed randomisation (12) and further uses of RAR in clinical trials to be severely limited (5).

In Section 2, we will look at the RAR rule used in the ECMO trial of (1) and discuss why it failed. In Section 3, we will introduce two dynamic programming-based rules as improvements and compare them using a range of metrics. We conclude in Section 4.

2 The ECMO Trial

The RAR rule used in the ECMO trial of (1) is the **randomised play-the-winner** (RPW) rule of (13). We assume there are one control group and one treatment group in our trial. The RPW rule is a randomised version of the **Play-The-Winner rule** by (15), which is a deterministic rule that will allocate the next patient based on the performance of the previous patient - if the previous patient is allocated to a particular group and had a successful/unsuccessful outcome, then the next patient will be allocated to the same/opposite group. The RPW rule is based on an urn model, where we start off with u balls for each of the control group and treatment group. Every time a new patient enters the trial, a ball is drawn with replacement to determine which group the patient is allocated to. The urn will then be updated based on the outcome of the patient - if the outcome is successful, β balls will be added to that group and α balls will be added to the opposite group. Here, $\beta \geq \alpha \geq 0$. This means we will start with an equal probability of allocation to control and treatment, but as there are more outcomes, the probability of allocation for the more successful group would be higher. Such an allocation rule is called an RPW(u, α, β) rule.

It is quite obvious that the specifications of the parameters u, α, β would induce very different allocation behaviours. For example, if the relative sizes of α and β are much smaller than u, the allocation probability would not be altered drastically for a small number of patient outcomes. On the other hand, if α and β are about the same or larger than u, the allocation probability would be greatly altered by only a few patient outcomes. This is exactly what happened during the ECMO trial of (1), where they used an RPW$(1, 0, 1)$ rule. Note that this is the same as the Thompson sampling with Beta-Bernoulli conjugacy and the Beta prior has parameters $(1, 1)$ which is exactly the value of u.

The control group (C) of the ECMO trial was treated with conventional therapy, and the treatment group (T) was treated with ECMO. The sample size calculations stated that if the survival rate gap is greater than 0.4, a sample size of 10 would give 95% of the patients to the superior treatment. The outcome of this trial is listed in the table below.

Allocation	T	C	T	T	T	T	T	T	T	T	T	T
Outcome	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Outcome of the ECMO trial of (1)

The clear imbalance of the allocations is the major criticism of this work, and the imbalance could be significantly remedied by a change in the parameterisation of the RPW rule. To investigate[1] the effect of u values on the final allocation ratio, we ran 100,000 Monte Carlo iterations with the true success probability of C and T being 0.2 and 0.65 respectively according to the previous evidence mentioned in (1). We fix $\alpha = 0, \beta = 1$, and vary u, while keeping the total number of patients to be 12.

u	$\mathbb{E}[\#C]$	$\mathbb{E}[\#T]$	$\mathbb{P}[\#T \geq 11]$
1	4.39	7.61	0.051
5	5.17	6.83	0.012
10	5.46	6.54	0.007

Table 2: Monte Carlo Simulation Results Summary

Note that the last column, to a certain extent, represents the one-sided p-value of observing the result of Table 1. The ratio is relatively balanced for $u = 1$, in fact, but it would be much better if a higher value, say $u = 5$, is used instead. That would also yield a more convincing argument for the ECMO trial. However, one should also keep in mind that the statistical significance of a trial with size 12 is ultimately limited.

3 Alternative RAR Rules and Performance Assessments

3.1 Optimal Design via Dynamic Programming

The response adaptive design problem with two arms and binary outcomes, such as the ECMO trail in Section 2, can be viewed as an MABP, which could be formulated formally as a dynamic programming (DP) problem and solved given some formal notion of rewards (8). This realisation allows us to design RAR rules more rigorously via finding the optimal policy of the dynamic programme, and two such policies/rules, studied in (14), are introduced in this and following subsections. Note that we adopt similar notations as that of (14). A formal treatment of DP for MABP can be found in Chapter 4 of (8), where concepts such as value function, value iteration, and policy improvements are studied.

We assume the patients arrive one by one, and the outcome is known immediately. The two treatment groups are denoted by A and B. The trial is assumed to be run for a fixed number of patients n. For each patient, the outcome of the treatment is modelled by a Bernoulli random variable, which we denote by $X \sim \text{Ber}(\theta_A)$ and $Y \sim \text{Ber}(\theta_A)$ for groups A and B with success probabilities θ_A and θ_B respectively. The reward is set to be the total expected number of successes of the whole trial.

Using a Bayesian approach, we would set priors on θ_A and θ_B, and as the likelihood is Bernoulli, we would use the conjugate prior which is the Beta distribution $\text{Beta}(\alpha, \beta)$. The following table summarises the update rules after knowing the outcomes of t patients.

[1] Code for the simulation can be found here.

	Prior (α, β)	# **Succ.** & **Fail. at** t	**Posterior** (α, β)
A	$(s_{A,0}, f_{A,0})$	$(s_{A,t}, f_{A,t})$	$(\tilde{s}_{A,t}, \tilde{f}_{A,t}) = (s_{A,0} + s_{A,t}, f_{A,0} + f_{A,t})$
B	$(s_{B,0}, f_{B,0})$	$(s_{B,t}, f_{B,t})$	$(\tilde{s}_{B,t}, \tilde{f}_{B,t}) = (s_{B,0} + s_{B,t}, f_{B,0} + f_{B,t})$

Table 3: Prior and Posterior Specifications

Therefore, the value function representing the total maximum expected reward after observing t results is

$$F_t(s_A, f_A, s_B, f_B)$$
$$= \max_{\pi \in \Pi} \mathbb{E}_\pi \left[\sum_{v=t}^{n-1} \sum_{g \in \{A,B} \frac{\tilde{s}_{g,v}}{\tilde{s}_{g,v} + \tilde{f}_{g,v}} \delta_{g,v} \middle| \tilde{s}_{A,t} = s_A, \tilde{f}_{A,t} = f_A, \tilde{s}_{B,t} = s_B, \tilde{f}_{B,t} = f_B \right]$$

where Π is the family of all allocation policies that assign one group to each patient and $\delta_{g,v}$ is the indicator variable of patient v for group g (e.g. if patient 8 is allocated group A, then $\delta_{A,8} = 1, \delta_{B,8} = 0$). This problem can be solved using standard dynamic programming techniques such as backward induction or the Whittle index (8), and the exact algorithm formulations, denoted by DP, can be found in Appendix A.1 of (14).

3.2 Optimal Design via Constrained Randomised DP

As noted in (11), the above algorithm is completely deterministic and will have low statistical power. Neither of the two properties is good, and therefore (14) proposed an alternative reward for the dynamic programming by adding randomness into the allocation rule, as well as constraining via penalising allocating fewer patients than a predetermined threshold. This algorithm is denoted as CRDP as it is derived by considering a constrained randomised dynamic program.

To be slightly more specific, when we choose to allocate a group A (or B) to a patient, the randomness in allocation will mean that the patient will be allocated to A (or B) with probability p and to B (or A) with probability $1 - p$. Note that $p = 1$ yields the design in Section 3.1. The added penalty in the reward is made to ensure that at least a predetermined m number of patients are allocated to each of the two groups. For the value function that represents the reward, a very large M would be deducted from it if the policy considered does not have m patients in each of the two groups, which would imply that no such policy would be considered in the dynamic programming. The full algorithm is omitted here due to the length constraint of this report, and interested readers are referred to Sections 2.2 and 2.3 of (14).

3.3 Performance Assessments and Numerical Comparisons

The main features of the ECMO trial can be summarised as: (1) the difference in allocated group sizes $|\#C - \#T|$ is too big, (2) the number of patients allocated to the better arm is high. The first point can be viewed as a criticism, while the second point

can be a compliment. Recall the ethical dilemma of a clinical trial, the first point shows a lack of collective benefit while the second point shows an emphasis on individual benefits. A more well-rounded assessment set of metrics is thus needed for better comparisons of allocation rules. In (5), the authors proposed a list of metric groups:

- Testing metrics (e.g. type 1 error, power)
- Estimation metrics (e.g. bias, variance, MSE)
- Patient benefit metrics (e.g. proportion of allocations to the best arm).

Here, we will compare the two algorithms of Sections 3.1 (DP) and 3.2 (CRDP), as well as the RCT using a few performance assessments for a more well-rounded comparison. The simulation[2] would be done on a trial with 100 patients, and the true success probabilities θ_A and θ_B of group A and B are set to be $(0.2, 0.7)$, $(0.4, 0.7)$, and $(0.6, 0.7)$. This is to consider a range of scenarios. The hypothesis test of the trial is $\theta_A - \theta_B = 0$ against its complement. The performance metrics are the test power of the algorithm using Fisher's exact test with p-value cutoff at 0.1, the MSE of the estimator for $\theta_A - \theta_B$ using sample proportions, and the proportion of allocation to the best (better) arm - each from one of the three groups of metrics above. The results obtained from 1000 Monte Carlo iterations are summarised in the following plots.

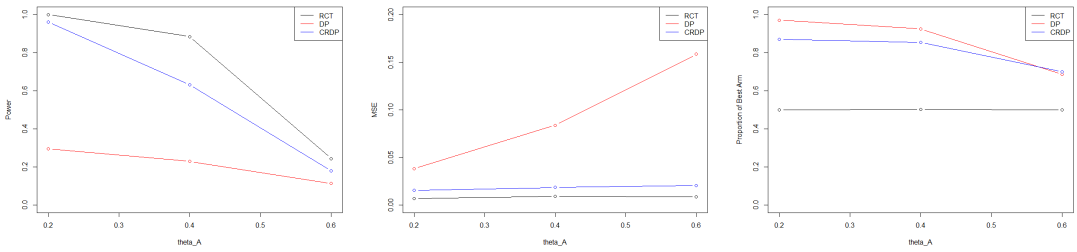

Figure 1: Simulation Results for Varying θ_A with 1000 Monte Carlo Iterations

It is clear that the RCT has good estimation and testing properties, but is bad at providing patient benefits. On the other hand, DP provides a significant amount of patient benefits, yet scores low on estimation and testing metrics. The CRDP sits right in the middle, as designed.

4 Summary

Overall, we have looked at a few response-adaptive randomisation (RAR) rules/algorithms and made critical evaluations and comparisons, as well as drawing their links to the multi-armed bandit problem. We have explored some of the main reasons why RAR rules are not used regularly in practice, and several further methodological works that aimed to develop better rules which could serve as the randomisation rules for certain

[2]Code for the simulation can be found here.

types of clinical trials, such as those for rare diseases where there are a limited amount of potential patients or those where failed treatments could be detrimental.

References

[1] R. H. Bartlett, D. W. Roloff, R. G. Cornell, A. F. Andrews, P. W. Dillon, and J. B. Zwischenberger. Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics*, 76(4):479–487, 1985.

[2] B. Freedman. Equipoise and the ethics of clinical research. *N Engl J Med*, 317(3-16):141–5, 1987.

[3] L. M. Friedman, C. D. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger. *Fundamentals of clinical trials*. Springer, 2015.

[4] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2):148–164, 1979.

[5] D. S. Robertson, K. M. Lee, B. C. López-Kolkovska, and S. S. Villar. Response-adaptive randomization in clinical trials. *Statistical Science*, 38(2):185, 2023.

[6] W. F. Rosenberger, O. Sverdlov, and F. Hu. Adaptive randomization for clinical trials. *Journal of biopharmaceutical statistics*, 22(4):719–736, 2012.

[7] D. Silver, A. Huang, C. J. Maddison, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[8] R. S. Sutton and A. G. Barto. *Reinforcement learning*. MIT Press, 2018.

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

[10] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[11] S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science*, 30(2):199, 2015.

[12] J. H. Ware. Investigating therapies of potentially great benefit: Ecmo. *Statistical Science*, 4(4):298–306, 1989.

[13] L. Wei and S. Durham. The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73(364):840–843, 1978.

[14] S. F. Williamson, P. Jacko, S. S. Villar, and T. Jaki. A bayesian adaptive design for clinical trials in rare diseases. *Computational Statistics Data Analysis*, 113:136–153, 2017.

[15] M. Zelen. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64(325):131–146, 1969.