Econometrics and Statistics 25 (2023) 125–133

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

A Markov decision process for response adaptive designs☆

Yanqing Yi[a,*], Xikui Wang[b]

[a] Division of Community Health & Humanities, Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL, Canada
[b] Warren Centre for Actuarial Studies and Research, Asper School of Business, University of Manitoba, Winnipeg, MB, Canada

ARTICLE INFO

Article history:
Received 22 March 2020
Revised 22 October 2021
Accepted 23 October 2021
Available online 1 November 2021

MSC:
62L05
62P10

Keywords:
Average reward criterion
Markov decision process
Randomization
Response adaptive design of clinical trials
Statistical power

ABSTRACT

The randomized treatment allocation process in a response adaptive clinical trial is formulated as a stochastic sequential decision problem and an algorithm is proposed to approximate the optimal value under the average reward criterion. When the information of previous treatment allocations and associated responses are summarized with sufficient statistics for unknown parameters, the decision process becomes a Markov process, on which a span-contractor operator is defined. It is proven that the average reward under the policy identified from the span-contractor operator converges almost surely to the optimal value. Numerical results reveal that the sequential procedure based on the controlled Markov process shows superior ethical advantage and at the same time produces good statistical power for large sample sizes such as 200 or larger.

1. Introduction

Randomized clinical trials are experiments on human subjects to assess new medical interventions and are typically characterized by the ethical tension between doing the best for collective benefits after the trials (i.e., collective ethics, or exploration) and treating effectively individuals in the trials (i.e., individual ethics, or exploitation). To fulfill the goal of collective ethics, the gold standard is to employ a balanced randomization in clinical trials. Consequently the statistical power is maximized and future patients may benefit. The use of balanced randomization is justified with clinical equipoise and it results in half of trial subjects being allocated to the inferior treatment. This is often criticized for being unethical in desperate situations because the clinical equipoise gradually loses its ground in the face of sequentially accumulating information and individual ethics is severely sacrificed (Pullman and Wang (2001)). Ethically motivated designs, including response adaptive designs (RADs), have been proposed to mitigate the ethical tension. The goal of RADs is to improve individual ethics while still safeguarding collective ethics and maintaining validity and integrity of the clinical research. In recent years, many RADs and statistical analysis methods have been proposed and investigated, including the play-the-winner design by Zelen (1969), drop-the-loser design (Ivanova (2003); Rosenberger and Hu (2004)), doubly adaptive biased coin design (Eisele (1994); Hu and Zhang (2004)), and covariate-adjusted response adaptive design (Hu et al. (2015)), just to name a few.

☆ The R code can be downloaded from the supplementary materials or from Github.

* Corresponding author.
E-mail address: Yanqing.Yi@med.mun.ca (Y. Yi).

https://doi.org/10.1016/j.ecosta.2021.10.015

RADs use the sequentially accumulated information during a trial and apply a skew randomization so that the potentially superior treatment is allocated to more trial subjects. Such an unbalanced and observations dependent randomization inevitably leads to unbalanced treatment group sizes on one hand and dependent samples on the other hand. Theoretically this results in a possible loss of statistical efficiency and traditional statistical analysis based on independent samples are invalid without modifications. The impact of treatment allocation variation of RADs on statistical power has been investigated by means of simulation (Melfi et al. (1998); Rosenberger et al. (2001); Korn and Freidlin (2011)) and in theory (Hu and Rosenberger (2003)). The specific design of unbalanced randomization in RADs often relies on some target optimal allocation proportions of alternative treatments in a trial, and various optimal allocation proportions have been proposed to achieve a certain goal of individual ethics while maintaining statistical power (Rosenberger et al. (2001); Tymofyeyev et al. (2007)). Recently, compound optimality criteria have been proposed to alleviate the tension between ethical benefits and the loss of statistical efficiency (Baldi Antognini and Giovagnoli (2010); Baldi Antognini et al. (2010); Baldi Antognini and Zagoraiou (2012); Hu et al. (2015); Metelkina and Pronzato (2017)). The target optimal proportions under various compound criteria and the optimal allocation proportions proposed by Rosenberger et al. (2001) are functions of unknown parameters of the distribution functions for patient responses. In practice, the judgment that one particular treatment is superior than other treatments are normally based on treatment effects beyond a given threshold. When designing a clinical trial, the target optimal proportions reduce to some minimum differences under the alternative hypothesis. Our objective in this paper is to explore adaptive methods of sequentially learning treatment effects to achieve the adaptive randomization of treatment groups at the desired proportions in order to detect the minimum differences.

In principle, adaptive random treatment allocation of the next patient in RADs depends on the summary statistic of treatment effects, based on the information of responses and treatments of previous patients in the trial. There are different approaches to summarize such information. Thompson (1933) adopted the Bayesian approach with a uniform prior and derived the probability of one treatment success probability higher than the other one. Wathen and Thall (2017) studied modified Bayesian designs of response allocation probability for multi-arm trials by combining three types of modifications to reduce undesired characteristics of the designs. Cheng and Berry (2007) used a Bayesian method for the unknown parameters and considered the optimality of restricted treatment allocations. Schmitz (1993) proved that the information process based on sequentially planned probability ratio tests is a homogeneous Markov process. Yi (2013) formulated the information gathering process as a Markov process using sufficient statistics for the unknown parameters and derived the exact type I error rate and statistical power for RADs with binary outcomes. Flournoy et al. (2013) proposed a graphical comparison method for RADs based on sufficient statistics. Ondra (2015) laid out a basic structure on the use of Markov decision programming for RADs under the discount reward criterion. In our paper, we consider the average reward criterion for RADs and propose an algorithm to approximate the optimal average reward and identify the optimal policy.

The remaining of our paper is organized as follows. Section 2 introduces the sequential decision model for RADs and the Markov policy. Section 3 discusses the approximation of the optimal values based on a span-contractor operator and introduces an algorithm for the sequential approximation. Section 4 conducts simulation studies using the algorithm. Section 5 concludes the paper.

2. The decision model for RADs and the Markov policy

This section introduces the sequential decision model for RADs and the allocation probability based on the sufficient statistics for the parameters of the response distributions.

We consider the comparison of two treatments A and B in this paper. The decision model can be generalized to trials with three or more treatments, but is not explored here. Suppose that trial subjects arrive sequentially and each subject receives one and only one of the two treatments. Responses $Y_{1j}, Y_{2j}, \cdots$ from treatment j are independent and identically distributed with a probability mass/density function $f_j(y, \theta_j)$ where $\theta_j \in \Theta_j$, $j = AorB$ is an unknown parameter; $y \in \mathbb{R}$, and $\mathcal{R}$ is the σ algebra over $\mathbb{R}$. Denote $\theta = (\theta_A, \theta_B)$. Our goal is to introduce a response adaptive design to achieve two simultaneous objectives: making a statistical comparison of alternative medical interventions at the conclusion of the trial and allocating as many trial subjects as possible to the potentially superior treatment.

Let δ_i be the treatment allocation for the i^{th} subject such that $\delta_i = 1$ if the i^{th} subject receives treatment A and $\delta_i = 0$ otherwise, and $\mathbf{y}_i = (Y_{iA}\delta_i, Y_{iB}(1-\delta_i))$ be the corresponding responses. Here we use the convention that if treatment j, $j = A$ or B, is not allocated to subject i, then the response is set to 0. When the i^{th} subject ($i > d$) is to be treated, the information available is given by the σ algebra $\mathcal{F}_{i-1}$ generated by $\{(\delta_1, \mathbf{y}_1), \cdots, (\delta_{i-1}, \mathbf{y}_{i-1})\}$, where d is an even integer, representing the number of subjects allocated using a non-adaptive randomization at the beginning of the trial.

A response adaptive design is characterized by a sequence of treatment allocation probabilities $\pi = \{\pi_i,\ i = 1, 2, \cdots\}$ such that each π_i is a stochastic kernel on $\mathbb{A}$ given $\{(\delta_1, \mathbf{y}_1), \cdots, (\delta_{i-1}, \mathbf{y}_{i-1})\}$, where $\mathbb{A} = \{0, 1\}$. That is, $\pi_i = P(\delta_i = 1 | (\delta_1, \mathbf{y}_1), \cdots, (\delta_{i-1}, \mathbf{y}_{i-1}))$, $i > d$, is a $\mathcal{F}_{i-1}$-measurable function. For the first d subjects, we prefix $\pi_i = P(\delta_i = 1)$ such as $1/2$, $i = 1, 2, \cdots, d$.

The sequential decision model for the response adaptive design is formulated as a tuple $((\mathbb{S}, \mathcal{S}), (\mathbb{A}, \mathcal{A}), \pi, (q_n), (\mathbf{y}_n))$, where

1. $\mathbb{S} = \mathbb{R}^2$ is the state space and its σ algebra $\mathcal{S}$ is the product of the σ algebra $\mathcal{R}$ on $\mathbb{R}$
2. $\mathbb{A}$ is the action space and its σ algebra $\mathcal{A}$ is the power set of $\mathbb{A}$.

3. $\pi = \{\pi_i,\ i = 1, 2, \cdots\}$ is the treatment allocation rule corresponding to the RAD in which π_i is a stochastic kernel from $(\mathbb{H}_{i-1}, \mathcal{F}_{i-1})$ to $(\mathbb{A}, \mathcal{A})$, $i > d$; $\pi_i = \pi_1, i = 1, 2, \cdots, d$ and π_1 is pre-defined based on prior knowledge on treatment effects; and $\mathbb{H}_n = \{(h, \delta, s), h \in \mathbb{H}_{n-1}, \delta \in \mathbb{A}, s \in \mathbb{S}\}, n > d$, where $\mathbb{H}_d = \{(\delta_1, s_1, \cdots, \delta_d, s_d) : \delta_i \in \mathbb{A}, s_i \in \mathbb{S}, i = 1, 2, \cdots, d\}$.
4. q_n is a transition probability law from $(\mathbb{H}_{n-1} \times \mathbb{A}, \mathcal{F}_{n-1} \times \mathcal{A})$ to $(\mathbb{S}, \mathcal{S})$, specified by treatment allocation δ_n and response distribution functions $f_j(y, \theta_j)$, $j = A$ or B, and is given by

$$q_n(\mathbf{y}_n \in C | h, \delta_n) = \left(\int_{C_1} f_A(y)dy\right)^{\delta_n} \left(\int_{C_2} f_B(y)dy\right)^{1-\delta_n},$$

where $h \in \mathbb{H}_{n-1}$, $\delta_n \in \mathbb{A}$, and C_l is the projection of C on the l^{th} component, $l = 1, 2$;
5. $r(\mathbf{y}_n, \delta_n)$ is a real-valued reward function on $\mathbb{S} \times \mathbb{A}$.

Let $\mathbb{H} = \mathbb{A} \times \mathbb{S} \times \mathbb{A} \times \mathbb{S} \times \cdots$, and $\mathcal{H}$ be its corresponding product σ-algebra. By Ionescu Tulcea's theorem, there is an unique probability measure P_π such that

$$P_\pi(\delta_n | h_n) = \pi_n^{\delta_n}(1 - \pi_n)^{1-\delta_n},$$
$$P_\pi(C | h_n, \delta_n) = \left(\int_{C_1} f_A(y)dy\right)^{\delta_n} \left(\int_{C_2} f_B(y)dy\right)^{1-\delta_n}.$$

for any C, δ_n and n. To simplify notation, let $\pi(\delta_n | h_{n-1}) = \pi_n^{\delta_n}(1 - \pi_n)^{1-\delta_n}$ in the rest of this paper.

The transition probability function of the stochastic process $\{(\delta_n, \mathbf{y}_n)\}$ under P_π is determined by the transition law $q_n((\delta_n, \mathbf{y}_n) | h_{n-1}) = \pi(\delta_n | h_n)[f_A(y_{iA}, \theta_A)]^{\delta_n}[f_B(y_{iB}, \theta_B)]^{1-\delta_n}$. The likelihood function of the observed sequence $\{(\delta_1, \mathbf{y}_1), \cdots, (\delta_n, \mathbf{y}_n)\}$ is given by

$$L(\theta) = \prod_{i=1}^{n} \left[\pi_n f_A(y_{iA}, \theta_A)\right]^{\delta_n} \left[(1 - \pi_n) f_B(y_{iB}, \theta_B)\right]^{1-\delta_n} = h(\pi) \prod_{i=1}^{n} \left[f_A(y_{iA}, \theta_A)\right]^{\delta_n} \left[f_B(y_{iB}, \theta_B)\right]^{1-\delta_n}$$

where $h(\pi) = \prod_{i=1}^{n} \pi_i^{\delta_i}(1 - \pi_i)^{1-\delta_i}$, and $0^0 = 1, \infty^0 = 1$.

For special response distribution functions such as the exponential distribution family, sufficient statistics can be found based on the likelihood function and they contain the information of treatment effects. Also, unconditional analysis approaches are advocated (Rosenberger and Lachin (2002); Proschan and Nason (2009)). When using the sufficient statistics to summarize information on treatment effects and employing the summarized information to determine the allocation probability π_i for the next patient, the treatment allocation process becomes a Markov decision process (Wei (1990); Yi (2013)).

The randomized treatment allocation rules based on the sufficient statistic are Markovian and include a wide range of response adaptive designs, including the RPW design, the drop-the-loser design (Ivanova (2003)), the optimal adaptive design proposed by Rosenberger et al. (2001), the generalized Pólya urn design (Wei (1979)), and the doubly adaptive biased coin design (Eisele (1994); Hu and Zhang (2004)). The treatment allocation process becomes a Markov decision process under the sequential decision model if its corresponding allocation rule π is Markovian.

3. The stationary policy and MDP procedure

In this section, we establish the optimality theorem and approximate the optimal value from the perspective of a Markov decision process (MDP). We then integrate the method of value iteration with the information process to sequentially identify the optimal policy. Although the information process, defined with sufficient statistics for the unknown parameters of the treatment effect, depends on the number of subjects receiving each treatment and their responses, the determination of treatment randomization probability for the next subject is based only on the size of observed treatment effect by means of the reward revealed by the information. That is to say, no matter when the anticipated magnitude of the treatment effect is observed, the treatment allocation probability for the next subject remains the same and constitutes a stationary policy.

To maintain randomness and provide learning on θ_j, $j = A, B$, we consider a policy $\pi = \{\pi_i,\ i = 1, 2, \cdots\}$ such that $\gamma \le \pi_i \le 1 - \gamma$, where γ is a constant between 0 and 0.5, and the choice of value of γ depends on the total size considered. Please see the discussion in Section 4 for further details. Cheng and Berry (2007); Wathen and Thall (2017) considered this type of policy and investigated its optimality from the Bayesian perspective. The simulation results from Wathen and Thall (2017) revealed that the design with 10 subjects each assigned to each of the treatments by complete randomization initially and then adaptively using π_i restricted between 0.1 and 0.9 has the desired properties of unbalanced group sizes for binary responses. Let $\Pi = \{\pi : \gamma \le \pi_i \le 1 - \gamma\}$.

We consider the average reward $V(\pi) = \liminf_{n\to\infty} E_\pi\left(\sum_{i=1}^{n} r(y_i, \delta_i)\right)/n$ for any policy π. Let $V^* = \max_{\pi\in\Pi} V(\pi)$ be the optimal value of RADs under the average reward criterion. Our goal is to find an optimal policy π^* such that $V(\pi^*) = V^*$. Such an optimal policy exists because the policy space Π is convex in that for any $\pi, \pi' \in \Pi$ and $0 < a < 1$, there exists $\pi'' \in \Pi$ such that $P_{\pi''} = aP_\pi + (1 - a)P_{\pi'}$.

We approximate the optimal value V^* using a span-contractor and propose a procedure based on a Markov decision process to identify the optimal policy π^*. For any $\pi \in \Pi$, Yi and Wang (2007) proved that $\frac{\sum_{i=1}^{n}\pi_i}{n} - \frac{N_A}{n} \to 0$ almost surely (a.s.) as $n \to \infty$, where $N_A = \sum_{i=1}^{n}\delta_i$. Hence, $\gamma \le N_A/n \le 1-\gamma$ and $N_A \xrightarrow{a.s.} \infty$. When N_A/n converges a.s. as $n \to \infty$, $V(\pi) = \lim_{n\to\infty} E_\pi\left(\frac{N_A}{n}r(A) + (1-\frac{N_A}{n})r(B)\right)$, where $r(A) = E(r(Y_A,\delta)|\delta = 1)$ and $r(B) = E(r(Y_B,\delta)|\delta = 0)$. Therefore, $V^* = (1-\gamma)\max\{r(A), r(B)\} + \gamma\min\{r(A), r(B)\}$. The reward function $r(Y,\delta)$ can be generic as a function of response Y. A simple example is $r(Y,\delta) = \delta Y_A + (1-\delta)Y_B$. Therefore $r(A) = E(Y_A)$ and $r(B) = E(Y_B)$.

Lemma 3.1. *For any given policy $\pi \in \Pi$, the transition probability $q(.|h)$ satisfies*

$$\sup_{h,h'} ||q(.|h) - q(.|h')|| \le 2|1-2\gamma|,$$

where $||\,.\,||$ is the variation norm.

For any $\delta \in \mathbb{A}$ under a policy π and any $C \in \mathbb{R}$, $||q(.|h) - q(.|h')|| = 2sup_{(\delta,C)}|q(\delta,C|h) - q(\delta,C|h')|$ and $q((\delta,C)|h) = \pi(\delta|h)P(Y_\delta \in C)$. The result of Lemma 3.1 is valid since $\gamma \le \pi(\delta|h) \le (1-\gamma)$.

The information contained in an observed history h can be summarized with a sufficient statistics and further by the observed size of treatment effect x, which is used to determine π. After the allocated treatment is observed, the subject's response depends on the corresponding probability distribution function of the allocated treatment. For any bounded measurable function $u(x)$, define $Tu(x)$ as

$$\max_{\gamma\le\pi\le 1-\gamma}\left[\pi(x)\left(r(A) + \int u(b_A(x,y))f_A(y)dy\right) + (1-\pi(x))\left(r(B) + \int u(b_B(x,y))f_B(y)dy\right)\right],$$

where $\pi(x) = P(\delta = 1|x)$ and $b_k(x,y), k = A, B$, are updated treatment effects after observing (δ, y).

From Lemma 3.1, we can prove the following result.

Lemma 3.2. *The operator T is a span-contraction operator. That is, for functions $u_1(x)$ and $u_2(x)$,*

$$sp(Tu_1 - Tu_2)(x) \le (1-2\gamma)sp(u_1 - u_2),$$

where $sp(u)$ is the span semi-norm of $u(x)$ defined by $sp(u) = sup_x u(x) - inf_x u(x)$.

The proof of Lemma 3.2 is given in the Appendix.

For any bounded function $u_0(x)$, let $u_n(x) = Tu_{n-1}(x), n = 1, 2, \cdots$. Suppose the decision π_n' at time n maximizes the operator $Tu_{n-1}(x)$. That is,

$$u_n(x) = \pi_n'(x)\left[r(A) + \int u_{n-1}(b_A(x,y))f_A(y)dy\right] + (1-\pi_n'(x))\left[r(B) + \int u_{n-1}(b_B(x,y))f_B(y)dy\right].$$

Define the policy $\pi' = \{\pi_n', n = 1, 2, \cdots,\}$. Using Lemma 3.2 and the span-fixed point theorem, we can prove the following result.

Theorem 3.3. *For the policy π' defined above, we have (1) π' is optimal, i.e. $V(\pi') = \max_{\pi\in\Pi} V(\pi)$*

(2) $\lim_{n\to\infty}\frac{\sum_{i=1}^{n} r(Y_i,\delta_i)}{n} = V^*, P_{\pi'} - a.s.$

(3) $\lim_{n\to\infty}\frac{N_A}{n} = \rho, P_{\pi'} - a.s.$, where $\rho = (1-\gamma)I_{\{\theta_A>\theta_B\}} + \gamma I_{\{\theta_A<\theta_B\}} + 1/2 I_{\{\theta_A=\theta_B\}}$ and I is the indicator function.

Theorem 3.3 implies that the average reward under the policy π' converges to the optimal value and the allocation proportions converge to the optimal proportion almost surely. The proof of these results are based on a locally stationary process $\tilde{\delta}_n$ with $P(\tilde{\delta}_n = 1) = \gamma$ to which the allocation process δ_n approximates under π'. The reason behind this is that the estimated treatment effect x based on the estimators of the parameters of the distributions of responses will be in the neighborhood of the true treatment effect because of the consistency property of the estimates (Yi and Wang (2007)). The results in Theorem 3.3 is proven using the fixed point theorem for the stationary process and the consistency of estimates.

We introduce a sequential procedure of adaptive treatment allocation based a Markov decision process (MDP) as follows. To reduce the effect of random error on the variability of treatment allocation, two tuning parameters ζ and η are introduced in the MDP procedure, where ζ is used for decision making during the iteration and η is employed for updating the iterated value of u_n. Suppose that the first d subjects are assigned using complete randomization (CR) to obtain information for allocation adaptation. This method is the same as the burn-in strategy used in Wathen and Thall (2017) in which $d = 50$ and 10 for each of the five treatments. Depending on the small sample properties of the sequential estimates used for adaptive allocation, d can be set as small as 2 when the estimates have good small sample properties. Otherwise, choose a large value of d. In Section 4 we use $d = 2$ for binary responses and $d = 10$ and 20 for the exponential distributions.

The MDP procedure is as follows:

Step 1. The first d (an even integer) subjects are randomized by $\pi_i = 1/2,\ i = 1, 2, \cdots, d$.

Step 2. Obtain $\hat{\theta}_A$ and $\hat{\theta}_B$ and determine

$$\pi_n = \begin{cases} 1-\gamma & \text{if } Av > Bv + \zeta, \\ \gamma & \text{if } Av < Bv - \zeta, \\ 1/2 & \text{if } |Av - Bv| \le \zeta. \end{cases}$$

Table 1
Simulated results and asymptotic statistical power for binary responses ($n = 100$)

		RSIHR		MDP ($\gamma = 0.35$)		MDP ($\gamma = 0.25$)		CR
θ_A	θ_B	$E\left(\frac{N_A}{n}\right)(s.d)$	power	$E\left(\frac{N_A}{n}\right)(s.d)$	power	$E\left(\frac{N_A}{n}\right)(s.d)$	power	power
.4	.4	.500(.031)	.049	.500(.093)	.047	.500(.142)	.049	0.05
.5	.4	.526(.029)	.251	.559(.088)	.240	.597(.132)	.230	0.263
.6	.4	.550(.027)	.633	.603(.071)	.606	.670(.099)	.572	0.654
.7	.4	.589(.026)	.922	.625(.057)	.903	.709(.069)	.872	0.935
.6	.6	.500(.021)	.051	.500(.100)	.047	.500(.157)	.049	0.05
.7	.6	.519(.019)	.284	.560(.090)	.258	.602(.138)	.249	0.277
.8	.6	.536(.018)	.718	.602(.071)	.683	.671(.097)	.649	0.723
.9	.6	.550(.017)	.974	.625(.057)	.968	.710(.067)	.951	0.980

where $j\nu = \hat{r}(j) + \int \hat{u}_{n-1}(b_j(x,y))\hat{f}_j(y)dy$, $j = A, B$; $\hat{u}_{n-1} = (n-1)[(1-\gamma)\hat{r}(A) + \gamma\hat{r}(B)]I_E + (n-1)[\gamma\hat{r}(A) + (1-\gamma)\hat{r}(B)]I_F + (n-1)[1/2\hat{r}(A) + 1/2\hat{r}(B)]I_G$ for $n > d$; $E = \{\hat{r}(A) > \hat{r}(B) + \eta\}$; $F = \{\hat{r}(A) < \hat{r}(B) - \eta\}$; and $G = \{|\hat{r}(A) - \hat{r}(B)| \le \eta\}$.

Step 3. According to π_n, collect information on (y_n, δ_n) and update $\hat{\theta}_A$ and $\hat{\theta}_B$. Go back to Step 2 until all subjects are treated.

4. Simulation studies

We apply the MDP procedure to conduct numeric studies for two types of responses - the binary responses and an example of continuous responses. We assume that treatment A is better and report the simulated average proportion of patients allocated to treatment A and the statistical power for different values of γ to explore the trade-off between the benefit in ethics and loss of statistical power. Adaptive treatment allocation results in unbalanced treatment group sizes, thus potential loss of statistical power. In the simulation, the test statistic for statistical power is the Wald-type statistic for both of binary responses and exponentially distributed responses. For a type I error rate of 0.05, the critical value for statistical power is 1.645 for binary responses and 1.82 for the exponential distributions. We consider total sizes of 100 and 200 in the simulation to examine the performance of the MDP procedure, the average reward under which is proven to converge to the optimal value as $n \to \infty$. For sample sizes smaller than 100, it is recommended to use the exact distribution method (Yi (2013)). All the results are based on 10^6 runs.

In the simulation, the reward function is $r(Y, \delta) = \delta Y_A + (1-\delta)Y_B$ and the tuning parameters are set as $\eta = 0.1$ and $\zeta = 0.05$. Large values of η and ζ have a tendency of treating treatments A and B as equivalent even if the two treatments are different, thus decreasing the average proportion of patients $E(N_A/n)$ allocated to the better treatment (Table 4 in the Appendix). For small values of η and ζ, the random error in responses may influence the decision making substantially thus leading to a probability $P(N_A/n < 0.5)$ or $P(N_A/n < 0.25)$ substantially larger than those for $\eta = 0.1$ and $\zeta = 0.05$ (Table 4 in the Appendix).

For binary responses, we consider two scenarios for treatment B: $\theta_B = 0.4$ and 0.6. Let $d = 2$ for binary responses. That is, the first two subjects are allocated by complete randomization and the rest follows the adaptive procedure. The adjusted method proposed by Agresti and Caffo (2000) is used to sequentially estimate θ_A and θ_B for allocation adaptation, where θ_A and θ_B are the success probabilities for treatments A and B, respectively. The estimates are $\hat{\theta}_A = \frac{s_A+1}{n_A+2}$ and $\hat{\theta}_B = \frac{s_B+1}{n_B+2}$, where n_j and s_j are the number of subjects allocated to and the number of successes for treatment j, $j = A, B$, respectively. The adjusted estimates have good small sample properties and it was used in Rosenberger et al. (2001). With this method, $d = 2$ produces type I error rates close to the nominal level of 0.05 for sample sizes of 100 and 200 (Tables 1 and (2) for the simulation scenarios considered.

We conduct simulation studies to compare this MDP procedure with the optimal proportion proposed by Rosenberger et al. (2001) (RSIHR) for binary responses. The RSIHR proportion is targeted by the doubly biased coin method (DBCD) with the allocation function $g(x, \rho)$ (Hu and Zhang (2004)) in which α is set to be 100 as in Hu et al. (2006). This RSIHR design is expected to have a small variation in allocation proportions from perspectives of optimality of the RSIHR proportion (Rosenberger et al. (2001)) as well as the DBCD procedure with $g(x, \rho)$ (Hu et al. (2006)). In our MDP procedure, we consider $\gamma = 0.35$ and 0.25. Theoretically, small values of γ implies good ethical benefits because a larger proportion of subjects is allocated to the potentially better treatment. However, the statistical power would be reduced due to the very unbalanced treatment group sizes and the variation introduced by the adaptation of treatment allocation (Melfi et al. (1998); Rosenberger et al. (2001); Hu and Rosenberger (2003)). For other values of γ, the statistical power for the MDP procedure can be found in Figure 1 in the Appendix, which shows substantial loss of statistical power for every small γ.

The simulated results are summarized in Tables 1 and for the total sample sizes of 100 and 200, respectively. The last column in the two tables is the asymptotic statistical power under CR with half of patients allocated to each of the two treatments. The results in Tables 1 and 2 show that both of the RSIHR and MDP allocate higher than 50% of subjects to the better treatment but the statistical powers under the two RADs are lower than those under CR. The unbalanced groups sizes under RSIHR and MDP lead to the loss in statistical power in the comparison with CR. The results on reduced statistical

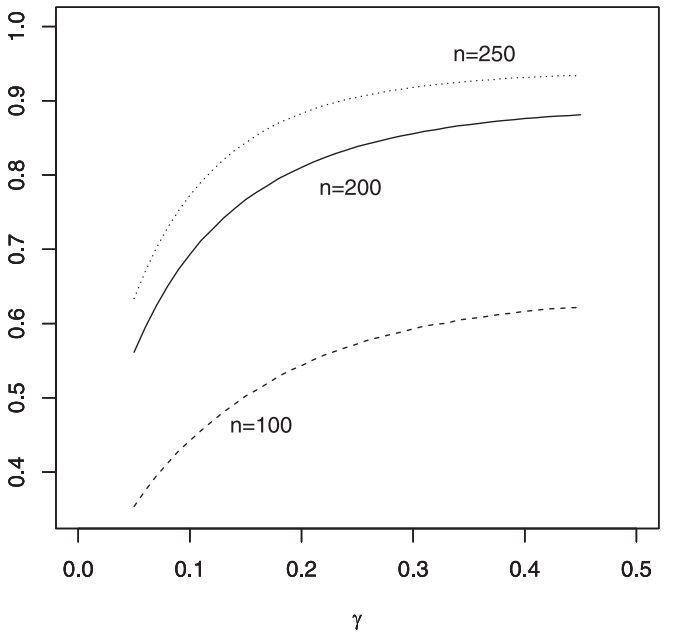

Fig. 1. Statistical power functions for binary responses ($\theta_A = 0.6, \theta_B = 0.4$) under MDP with $\gamma = 0.25$

Table 2
Simulated results and asymptotic statistical power for binary responses ($n = 200$)

		RSIHR		MDP ($\gamma = 0.35$)		MDP ($\gamma = 0.25$)		CR
θ_A	θ_B	$E\left(\frac{N_A}{n}\right)(s.d)$	Power	$E\left(\frac{N_A}{n}\right)(s.d)$	Power	$E\left(\frac{N_A}{n}\right)(s.d)$	Power	Power
.4	.4	.500(.022)	.049	.500(.082)	.049	.500(.134)	.052	0.05
.5	.4	.528(.020)	.407	.576(.071)	.395	.627(.111)	.379	0.414
.6	.4	.550(.019)	.883	.622(.048)	.868	.702(.066)	.838	0.893
.7	.4	.569(.018)	.996	.637(.038)	.995	.729(.042)	.990	0.998
.6	.6	.500(.015)	.048	.500(.089)	.049	.500(.147)	.052	0.05
.7	.6	.519(.014)	.430	.580(.072)	.420	.634(.115)	.402	0.439
.8	.6	.536(.013)	.928	.620(.047)	.910	.705(.064)	.898	0.935
.9	.6	.550(.012)	1.00	.637(.037)	1.00	.729(.041)	.999	1.00

power under response adaptive allocations align with those in Korn and Freidlin (2011), who reported increasing sample sizes under adaptive randomization over CR to maintain the same statistical power. Although the randomization in RADs is not expected to create imbalance in other covariates of subjects, imbalance could occur among treatment groups if those covariates in the study population change over time, thus leading to possible bias. In this case, blocked randomization or covariates adjusted RADs should be used (Hu et al. (2015)).

Tables 1 and 2 show the effect of the magnitude of γ on the trade-off between ethical benefits and the loss of statistical power. The MDP with $\gamma = 0.25$ allocates a higher proportion of subjects, on average, to the better treatment than, but the statistical power is lower than the MDP with $\gamma = 0.35$. While comparing with CD, the loss of statistical power under the MPD procedure is substantial for both $\gamma = 0.35$ and 0.25 when the total sample size is 100, where big losses are observed for the scenarios with $\theta_A - \theta_B = 0.2$ (Table 1) and (the biggest loss reaches 0.082 for $\theta_A = 0.6$ and $\gamma = 0.25$. However, when the total size increases to 200, the loss of statistical power decreases and the statistical power becomes close to the ones under CR for $\theta_A - \theta_B = 0.3$. In the comparison with the RSIHR design, the average proportion of patients allocated to the better treatment is higher under the MDP procedure but the statistical power is smaller. This indicates the trade-off between the skewed allocation and the loss of statistical power. In any case, the statistical power under the MDP procedure is quite close to the power under the RSIHR design and CR for a sample size as large as 200 and a large difference to be detected, say 0.3, between the two treatments. Finally on average, at least 6.8% and 16% more subjects are treated with the better treatment with the MDP procedure than with the RSIHR design.

We consider an exponential distribution of responses as a continuous case for the scenario $\theta_B = 5$ and assume that a treatment with a large value of mean responses is better. Suppose that all responses of treated subjects are available when the next patient is to be allocated. That is, we do not consider delayed responses. Please refer to Bai et al. (2002) and Hu et al. (2008) for models with delayed responses.

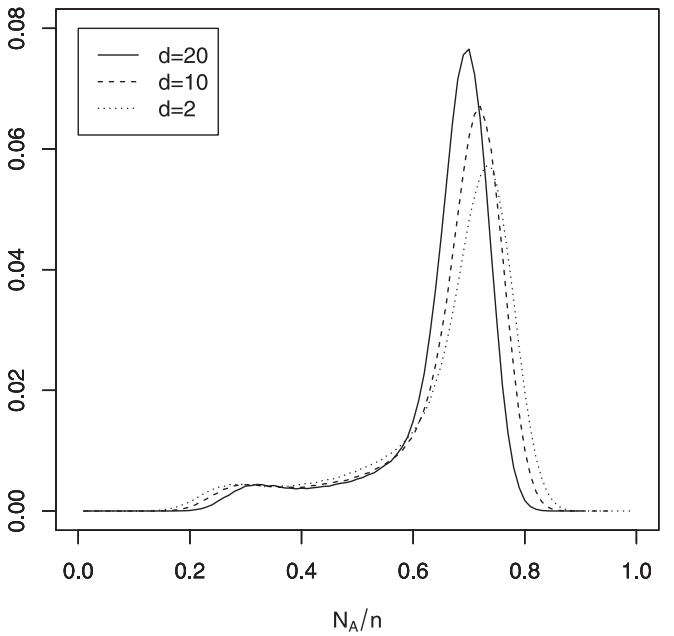

Fig. 2. Distribution of N_A/n for exponentially distributed responses ($\mu_A = 7$, $\mu_B = 5$, n=100) under MDP with $\gamma = 0.25$

Table 3
Simulated results for exponentially distributed responses (n=100)

		d=10			d=20			CR
θ_A	θ_B	$E\left(\frac{N_A}{n}\right)(s.d)$	Power	$P\left(\frac{N_A}{n} < 0.5\right)$	$E\left(\frac{N_A}{n}\right)(s.d)$	Power	$P\left(\frac{N_A}{n} < 0.5\right)$	Power
5	5	0.500(0.183)	0.051	0.495	0.500(0.166)	0.049	0.499	0.050
6	5	0.598(0.162)	0.198	0.252	0.594(0.145)	0.197	0.238	0.230
7	5	0.657(0.125)	0.435	0.114	0.649(0.107)	0.440	0.098	0.500
8	5	0.687(0.093)	0.671	0.050	0.676(0.077)	0.678	0.039	0.727
9	5	0.703(0.072)	0.836	0.023	0.688(0.058)	0.843	0.015	0.865

Table 4
Simulated results for values of η and ζ for binary responses ($p_A = 0.6$, $p_B = 0.4$, $n = 200$) under MDP with $\gamma = 0.25$

	$\eta = 0.1$			$\zeta = 0.05$	
	$\zeta = 0.025$	$\zeta = 0.05$	$\zeta = 0.1$	$\eta = 0.05$	$\eta = 0.15$
$E\left(\frac{N_A}{n}\right)$	0.706	0.702	0.609	0.708	0.699
s.d	0.065	0.066	0.078	0.069	0.064
Power	0.833	0.838	0.853	0.828	0.835
$P\left(\frac{N_A}{n} < 0.5\right)$	0.0163	0.0167	0.0324	0.0201	0.0160
$P\left(\frac{N_A}{n} < 0.25\right)$	1.7×10^{-4}	6.5×10^{-5}	1.6×10^{-5}	1.7×10^{-4}	8.0×10^{-5}

For the continuous responses, the magnitude of d has substantial impact on the distribution of allocation proportions when the maximum likelihood estimators of θs are sequentially used for adaptation. In the simulation scenarios, small values of d produce distributions of N_A/n with larger variation and thicker left tails while large values of d tend to shift the distributions to the left. See Figure Appendix A in the Appendix for $d = 2, 10,$ and 20. A thick left tail of the distribution of N_A/n indicates large $P(N_A/n < 0.5)$, implying a large chance that the trial goes to the wrong direction while treatment A is set as the better treatment. Therefore, a small value of $P(N_A/n < 0.5)$ is preferred. We report the simulated results for $d = 10$ and $d = 20$ in Table 3 for the total size of $n = 100$.

In the MDP procedure for the exponential distributions, $\gamma = 0.25$, $\eta = 0.1$ and $\zeta = 0.05$. Those values of η and ζ are the same as those for binary responses and they work well for the exponential distributions as well. For other distributions of responses, it is recommended to use those values as references to search for potentially better values of the tuning parameters. When $d = 20$, 10 subjects each are allocated to treatments A and B, respectively, and only 80 subjects are assigned by the MDP procedure in contrast to 90 such subjects under $d = 10$. Table 3 reveals that the MDP procedure with $d = 20$ produces smaller $E(N_A/n)$ with smaller standard deviation of N_A/n and a higher statistical power than those under $d = 10$. When θ_A is close to θ_B, $P(N_A/n < 0.5)$ is close to 1/4 but it decreases to 0.05 for $d = 10$ and to 0.039 for $d = 20$

as the difference between θ_A and θ_B increases to 3. Table 3 also shows that the difference in $P(N_A/n < 0.5)$ is substantial between $d = 20$ and $d = 10$ and so is the loss in statistical power of MDP in comparison with CR. The loss in statistical power is observed to decrease to 0.022 and 0.029 for $d = 20$ and $d = 10$, respectively, when the difference between θ_A and θ_B increases to 4. Therefore, the MDP procedure is recommended for the comparison of exponential distributions with moderate or large effect sizes and setting d as large as or larger than 20.

5. Conclusion

This paper introduces a sequential decision model for response adaptive clinical trials and considers restrictive adaptive randomization, for which a span-contraction operator is established. When sufficient statistics for the unknown parameters are used to summarize the information gathered on the treatment effect, the decision process becomes Markovian and a sequential procedure for response adaptive designs is proposed. The optimal value under the average reward criterion can be approximated through the value iteration based on a span-contractor. Simulation studies are conducted, which show ethical advantages of our proposed procedure for the scenarios considered. The statistical power of the proposed method remains reasonably good for large sample sizes to detect large differences between the two treatments.

Acknowledgement

The authors thank the anonymous associate editor and referees for their constructive criticisms and very helpful comments. The authors acknowledge research support from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Appendix A. Proof of Lemma 3.2

Assume that $\pi^{1*}(x')$ and $\pi^{2*}(x)$ maximize $Tu_1(x')$ and $Tu_2(x)$, respectively, where

$$Tu_k(x) = \max_{\gamma \le \pi \le 1-\gamma} \left[\pi(x)\left(r(A) + \int u_k(b_A(x,y))f_A(y)dy\right) + (1-\pi(x))\left(r(B) + \int u_k(b_B(x,y))f_B(y)dy\right)\right],$$

where $k = 1, 2$. Then,

$$\begin{aligned}(Tu_1 - Tu_2)(x') \le\ & \pi^{1*}(x')\left(r(A) + \int u_1(b_A(x',y))f_A(y)dy\right)\\ & +(1-\pi^{1*}(x'))\left(r(B) + \int u_1(b_B(x',y))f_B(y)dy\right)\\ & -\pi^{1*}(x')\left(r(A) + \int u_2(b_A(x',y))f_A(y)dy\right)\\ & -(1-\pi^{1*}(x'))\left(r(B) + \int u_2(b_B(x',y))f_B(y)dy\right)\\ =\ & \int (u_1 - u_2)q^{\pi^{1*}}(dy|x'),\end{aligned}$$

where $q^{\pi^{1*}}(E|x') = \pi^{1*}(x')\int I_{b_A(x',y)\in E}f_A(y)dy + (1-\pi^{1*}(x'))\int I_{b_B(x',y)\in E}f_B(y)dy$.

Similarly,

$$\begin{aligned}(Tu_1 - Tu_2)(x) \ge\ & \pi^{2*}(x)\left(r(A) + \int u_1(b_A(x,y))f_A(y)dy\right)\\ & +(1-\pi^{2*}(x))\left(r(B) + \int u_1(b_B(x,y))f_B(y)dy\right)\\ & -\pi^{2*}(x)\left(r(A) + \int u_2(b_A(x,y))f_A(y)dy\right)\\ & -(1-\pi^{2*}(x))\left(r(B) + \int u_2(b_B(x,y))f_B(y)dy\right)\\ =\ & \int (u_1 - u_2)q^{\pi^{2*}}(dy|x),\end{aligned}$$

where $q^{\pi^{2*}}(E|x) = \pi^{2*}(x)\int I_{b_A(x,y)\in E}f_A(y)dy + (1-\pi^{2*}(x))\int I_{b_B(x',y)\in E}f_B(y)dy$.

Define $\lambda(\cdot) = q^{\pi^{1*}}(\cdot|x') - q^{\pi^{2*}}(\cdot|x)$. Applying the Jordan-Hahn Decomposition Theorem and following the same steps in Hernández-Lerma (page 60), we have $\int (u_1 - u_2)\lambda(dy) \le (1-2\gamma)sp(u_1 - u_2)$. Therefore,

$$(Tu_1 - Tu_2)(x') - (Tu_1 - Tu_2)(x) \le (1-2\gamma)sp(u_1 - u_2).$$

By the arbitrariness of treatment effects x' and x, Lemma 3.2 is proved.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecosta.2021.10.015

References

Agresti, A., Caffo, B., 2000. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. American Statistician 54, 280–288.

Bai, Z.D., Hu, F., Rosenberger, W.F., 2002. Asymptotic properties of adaptive designs with delayed response. Annuals of Statistics 30 (1), 122–139.

Baldi Antognini, A., Giovagnoli, A., 2010. Compound optimal allocation for individual and collective ethics in binary clinical trials. Biometrika 97 (4), 935–946.

Baldi Antognini, A., Zagoraiou, M., 2012. Multi-objective of optimal designs in comparative clinical trials with covariates: the reinforced doubly adaptive biased coin design. The annuals of statistics 40 (3), 1315–1345.

Baldi Antognini, A., Zagoraiou, M., Giovagnoli, A., Atkinson, A., Torsney, B., coed, C.M., 2010. Covariate adjusted designs for combining efficiency, ethics and randomness in normal response trials, in. In: mODa9 - advances in model oriented design and analysis. Heidelberg: Physica-Verlag, Springer, Germany. 17-24, ISBN: 978-3-7908-2409-4

Cheng, Y., Berry, D.A., 2007. Optimal adaptive randomized designs for clinical trials. Biometrika 94, 673–689.

Eisele, J.R., 1994. The doubly adaptive biased coin design for sequential clinical trials. Journal of Statistical Planning and Inference 38, 249–261.

Flournoy, N., Haines, L.M., Rosenberger, W.F., 2013. A graphical comparison of response-adaptive randomization procedures. Statistics in Biopharmaceutical Research 5 (2), 126–141.

Hernández-Lerma, O., 2012. Discrete-time Markov control processes: basic optimality criteria. Springer.

Hu, F., Rosenberger, W.F., 2003. Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. Journal of the American Statistical Association 98, 671–678.

Hu, F., Rosenberger, W.F., Zhang, L.X., 2006. Asymptotically best response-adaptive randomization procedures. Journal of Statistical Planning and Inference 136, 1911–1922.

Hu, F., Zhang, L.X., 2004. Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. Annals of Statistics 32, 268–301.

Hu, F., Zhang, L.-X., Cheung, S.H., Chan, W.S., 2008. Doubly adaptive biased coin designs with delayed responses. The Canadian Journal of Statistics 36 (4), 541–559.

Hu, J., Zhu, H., Hu, F., 2015. A unified family of covariate-adjusted response-adaptive designs based on efficiency and ethics. Journal of the American Statistical Association 110 (509), 357–367.

Ivanova, A., 2003. A play-the-winner type urn model with reduced variability. Metrika 58, 1–13.

Korn, E.L., Freidlin, B., 2011. Outcome-adaptive randomization: Is it useful? Journal of Clinical Oncology 29 (6), 771–776.

Melfi, V.F., Page, C., Folurnoy, N., Rosenberger, W.F., Wong, W.K., 1998. Variability in adaptive designs for estimation of success probabilities. In: In new developments and applications in experimental design. Institute of Mathematical Statistics, Hayward, CA, pp. 106–114.

Metelkina, A., Pronzato, L., 2017. Information-regret compromise in covariate-adaptive treatment allocation. The Annals of Statistics 45 (5), 2046–2073.

Ondra, T., 2015. Optimized response-adaptive clinical trials - sequential treatment allocation based on Markov decision problems. Springer Spektrum.

Proschan, M.A., Nason, M., 2009. Conditioning in 2×2 tables. Biometrics 65, 316–322.

Pullman, D., Wang, X., 2001. Adaptive designs, informed consent, and the ethics of research. Controlled clinical Trials 22, 203–210.

Rosenberger, W.F., Hu, F., 2004. Maximizing power and minimizing treatment failures in clinical trials. Clinical Trials 1, 141–147.

Rosenberger, W.F., Lachin, J.M., 2002. Randomization in clinical trials: theory and practice. John Wiley and Sons, Inc., New York.

Rosenberger, W.F., Stallard, N., Ivanova, A., Harper, C.N., Ricks, M.L., 2001. Optimal adaptive designs for binary response trials. Biometrics 57, 909–913.

Schmitz, N., 1993. Optimal sequentially planned decision procedures. Springer-verlag, New York, Inc..

Thompson, W.R., 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of the two samples. Biometrika 25, 285–294.

Tymofyeyev, Y., Rosenberger, W.F., Hu, F., 2007. Implementing optimal allocation in sequential binary response experiments. Journal of the American Statistical Association 102 (477), 224–234.

Wathen, J.K., Thall, P.F., 2017. A simulation study of outcome adaptive randomization in multi-arm clinical trials. Clinical Trials 14 (5), 432–440. doi:10.1177/1740774517692302.

Wei, L.J., 1979. The generalized polya's urn design for sequential experiments. Annals of Statistics 7, 291–296.

Wei, L.J., 1990. Comments on 'on inferences from wei's biased coin design for clinical trials' by c. begg. Biometrika 77, 476–477.

Yi, Y., 2013. Exact statistical power for response adaptive designs. Computational Statistics and Data Analysis 58, 201–209.

Yi, Y., Wang, X., 2007. Goodness-of-fit test for response adaptive designs. Statistics and Probability Letters 77, 1014–1020.

Zelen, M., 1969. Play the winner rule and the controlled clinical trial. Journal of the American Statistical Association 64, 131–146.