Doubly adaptive biased coin designs with delayed responses

Feifang HU, Li-Xin ZHANG, Siu H. CHEUNG and Wai S. CHAN

Key words and phrases: Adaptive design; asymptotic normality; biased coin design; clinical trial; response-adaptive design; strong consistency; variance estimation.

MSC 2000: Primary 62E20; secondary 62L05.

Abstract: In clinical studies, patients are usually accrued sequentially. Response-adaptive designs are then useful tools for assigning treatments to incoming patients as a function of the treatment responses observed thus far. In this regard, doubly adaptive biased coin designs have advantageous properties under the assumption that their responses can be obtained immediately after testing. However, it is a common occurrence that responses are observed only after a certain period of time. The authors examine the effect of delayed responses on doubly adaptive biased coin designs and derive some of their asymptotic properties. It turns out that these designs are relatively insensitive to delayed responses under widely satisfied conditions. This is illustrated with a simulation study.

1. INTRODUCTION

1.1. Adaptive design.

In most clinical trials, patients are accrued sequentially. Response-adaptive designs are valuable and ethical randomization schemes that formulate treatment allocation as a function of the previous responses of patients. A major intention of such designs is to skew the probability of treatment allocation to increase the chance that more patients receive better treatment. The early important research work on adaptive designs can be traced back to Thompson (1933) and Robbins (1952).

In the literature, response-adaptive designs can be categorized into two classes. The first is the target-driven response-adaptive design, which is constructed using an optimal (or desired) allocation target in which a specific criterion is optimized based on a population response model. The second class is the design-driven response-adaptive design. The allocation rule for a member of this class is typically established with an intuitive motivation, but not with an optimal criterion in the formal sense (Rosenberger & Lachin 2002). A well-studied family of design-driven response-adaptive designs is based on urn models. The early influential developments in this area include the play-the-winner rule (for the comparison of two treatments with binary responses) of Zelen (1969) and the randomized play-the-winner rule of Wei & Durham (1978) and Wei (1979). An extensive review of the properties of urn models is provided in Rosenberger (2002), and some of the general asymptotic properties of urn models can be found in Janson (2004) and Bai & Hu (1999, 2005).
In target-driven response-adaptive designs, the doubly adaptive biased coin design (DBCD), which was first proposed in Eisele (1994) and Eisele & Woodroffe (1995), has recently gained recognition as a useful tool in developing treatment allocation schemes. The basic principles of the DBCD are inherited from those of Efron’s biased coin design (Efron 1971). However, the biased coin design depends only on the current proportion of subjects that is being assigned to each treatment, whereas the DBCD also depends on the current estimate of the desired allocation proportion.

The DBCD and its asymptotic properties are provided in Hu & Zhang (2004a) for multi-treatment clinical trials. Recent studies by Hu & Rosenberger (2003), Rosenberger & Hu (2004), and Hu & Zhang (2004a) demonstrate the important advantages of the DBCD, which can be summarized as follows. First, depending on the various possible desired treatment allocation proportions, the DBCD can be constructed to target prespecified allocation proportions, such as the urn allocation, the Neyman allocation, and the allocation of Rosenberger, Stallard, Ivanova, Harper & Ricks (2001). Second, compared to other adaptive designs with the same limiting allocation proportions, the DBCD generates smaller asymptotic variance, which means that it offers a higher power for statistical inferences, such as the test of the difference between treatment proportions for a fixed allocation (see Hu & Rosenberger 2003). Finally, the DBCD is very flexible and can be employed for general (continuous or discrete) responses (Hu & Rosenberger 2006; Tymofyeyev, Rosenberger & Hu 2007).

1.2. Delayed responses.

The DBCD has been demonstrated to be desirable for clinical trials with immediate outcomes, but in many circumstances individual patient outcomes may not be available before the randomization of the next patient, even though many outcomes are available during the recruitment period.

In terms of the treatment allocation mechanism, delayed responses pose no logistical difficulty in practice. In urn models, the urn can be updated when responses become available (Wei 1988), and in the DBCD, the estimates can be updated when the outcomes become available. Our main concern is how delayed responses affect response-adaptive designs, including the behavior of the allocation proportions and their variances. The effects of delayed responses in urn models were studied using simulations and reported in Rosenberger & Seshaiyer (1997). Recently, Bai, Hu & Rosenberger (2002) and Hu & Zhang (2004b) examined the effects of delayed responses for urn models theoretically, and they successfully proved that the asymptotic results remain unchanged by staggered entries with delayed responses under certain reasonable conditions.

It is still unclear how the delayed responses in the DBCD affect the estimators of the unknown parameters that may considerably influence other properties of the design. This is an important issue, as delayed responses are common phenomena in clinical studies, and, with the DBCD being a superior choice in adaptive designs, a clearer understanding of its qualities with respect to delayed responses will serve as a useful guideline to help practitioners choose from among the various treatment allocation schemes in clinical studies.

1.3. Objectives and organization of the paper.

The main purpose of this paper is to study the effects of delayed responses in DBCDs. We show that the asymptotic properties of the allocation proportions are unaffected by staggered entry and delayed responses in reasonable probability models. The small sample effects of delayed responses are evaluated through simulation studies.

The asymptotic normality that was obtained by Hu & Zhang (2004a) makes statistical inference possible with the DBCD without delayed responses. However, in practice, the asymptotic variance is unknown, and thus a good estimator of this variance is important for making accurate statistical inferences (Hu & Rosenberger 2003). In this paper, we also consider variance estimators for DBCDs with delayed responses. Strong consistency is obtained with these estimators.
The remainder of this paper is organized as follows. In Section 2, we start by proposing a treatment allocation scheme that incorporates delayed responses into the DBCD, and we supply some necessary notation and conditions. In Section 3, important asymptotic properties are studied and discussed. A simulation study is presented in Section 4 to evaluate the impact of delayed responses for different sample sizes. Some concluding remarks are given in Section 5. The technical proofs are provided in the Appendix.

2. DOUBLY ADAPTIVE BIASED COIN DESIGN WITH DELAYED RESPONSES

2.1. Notation and assumptions.

Suppose that in a clinical study, there are $K$ treatments, and the subjects arrive sequentially. Each patient is assigned to one of the $K$ treatments. Let $X_m = (X_{m,1}, \ldots, X_{m,K})$ denote the treatment assignment of patient $m$. If treatment $k$ is given to this patient, then $X_{m,k} = 1$, and the remaining entries of $X_m$ are all zero. For example, if $K = 4$ and the second patient receives the third treatment, then $X_2 = (0, 0, 1, 0)$.

Now, let the observed response of patient $m$ after treatment be $\xi_{m,k}$. For simplicity, we assume hereafter that the responses are univariate random variables, even though our results in this paper can be easily generalized to multivariate responses. For ease of the subsequent mathematical development, let $\xi_m = (\xi_{m,1}, \ldots, \xi_{m,K})$ represent the collection of possible responses of a patient with various treatments, even though only one particular treatment response, $\xi_{m,k}$, is actually observed. Assume that $\xi_{m,k}$, $m = 1, 2, \ldots$ are independent and identically distributed (i.i.d.) random vectors. Note that the responses of patients may not be available immediately after treatment.

After patient $m$ has been treated, let $N_m = (N_{m,1}, \ldots, N_{m,K})$, where $N_{m,k} = \sum_{j=1}^m X_{j,k}$ is the number of patients that have been assigned to treatment $k$, $k = 1, \ldots, K$, thus far. In addition, let $n$ be the total number of patients being treated at the end of the trial.

2.2. The treatment allocation scheme.

Let the probability of assigning treatment $k$ to the patient $(m+1)$ be $p_{m+1,k}$, $k = 1, \ldots, K$. The treatment allocation probability $p_{m+1,k}$ depends on

(a) the history of treatment assignment $N_m$ for the previous $m$ patients and

(b) the responses that have been observed thus far (there is a positive probability that the total number of responses observed up to this time is less than $m$ due to delayed responses).

Let the mean response of patients who are receiving treatment $k$ be $\theta_k = E\xi_{1,k}$, $k = 1, \ldots, K$. If all of the responses can be observed immediately, then to estimate $\theta_k$, the modified sample mean

$$\hat{\theta}_{m,k} = \frac{\sum_{j=1}^m X_{j,k}\xi_{j,k} + \theta_{0,k}}{N_{m,k} + 1} \tag{1}$$

can be used. Here, $\theta_{0,k}$ is the initial estimate (for dichotomous outcomes, .5 can be used if no prior knowledge is available about the treatment efficacy) of treatment $k$ when no patient has yet been assigned to it. Furthermore, 1 is added to the denominator to avoid a zero denominator. This minor adjustment plays a role only in the early stages of the clinical trial when the number of patients treated is still small. As more patients undergo treatment, $\hat{\theta}_{m,k}$ will become almost identical to the sample mean $\sum_{j=1}^m X_{j,k}\xi_{j,k}/N_{m,k}$.

As some of the responses may not be observed immediately, we let $T_{m,k} \leq N_{m,k}$, $k = 1, \ldots, K$ be the number of the responses that are observed up to the time when patient $(m+1)$ is ready for treatment assignment. Let the corresponding sum of the responses be $S_{m,k}$, $k = 1, \ldots, K$. The estimate of treatment efficacy (1) is modified to

$$\hat{\theta}_{m,k} = \frac{S_{m,k} + \theta_{0,k}}{T_{m,k} + 1}. \tag{2}$$
With the above updated estimates of treatment effects, we can now proceed to state the treatment allocation scheme.

The goal of the allocation scheme is to have \( N_n / n \) converge to a desired allocation proportion as \( n \to \infty \). In general, the desired allocation proportion is a function of the parameter vector \( \Theta = (\theta_1, \ldots, \theta_K) \). Denote the desired allocation proportion as \( v = \rho(\Theta) \), where \( \rho(\cdot) = (\rho_1(\cdot), \ldots, \rho_K(\cdot)) \) is a vector-valued function. Note that \( \rho(\cdot)^t = 1 \). That is, \( \rho_1(\cdot) + \cdots + \rho_K(\cdot) = 1 \) because each patient must receive one of the treatments. We have the following Treatment allocation scheme:

(a) Due to a lack of information about treatment efficacy in the initial stage, the first \( Km_0 \) patients are allocated to \( K \) treatments by using restricted randomization (see Rosenberger & Lachin 2002).

(b) For \( m \geq Km_0 \), patient \( (m + 1) \) is assigned to treatment \( k \) with a probability \( p_{m+1,k} = g_k(\frac{N_m}{m}, \hat{\rho}_m) \), \( k = 1, \ldots, K \), where

\[
\hat{\rho}_m = \rho(\hat{\Theta}_m),
\]

and \( \hat{\Theta}_m = (\hat{\theta}_m, 1, \ldots, \hat{\theta}_m, K) \) is the most recently updated estimate of \( \Theta \).

The treatment allocation functions \( g_k \), \( k = 1, \ldots, K \) depend on the history of treatment allocation \( N_m / m \) and the updated estimate of \( \Theta \). In general, the allocation functions are defined such that if the current sample allocation proportion \( N_{m,k} / m \) for a particular treatment \( k \) exceeds the estimated target \( \hat{\rho}_{m,k} \), then the next patient is assigned to this treatment with a probability that is smaller than \( \hat{\rho}_{m,k} \). For the choice of the allocation functions, refer to the discussions in Hu & Zhang (2004a).

### 2.3. Remarks on the response delay mechanism.

Before discussing the asymptotic properties, it is necessary for us to clearly describe the response delay mechanism. Let \( t_m \) be the entry time of patient \( m \), where \( t_m \) is an increasing sequence of random variables. The response time to treatment \( k \) of patient \( m \) with response \( \xi_{m,k} \) is denoted by \( \tau_{m,k} \), \( k = 1, \ldots, K \). Let \( M_k(m,l) = I\{t_m + \tau_{m,k} \in (t_{m+l}, t_{m+l+1})\} \) be an indicator function that takes the value 1 if the response of patient \( m \) on treatment \( k \), \( \xi_{m,k} \), is observed between the assignment of the treatments of patient \( (m+l) \) and patient \( (m+l+1) \). Obviously, for every pair of \( m \) and \( k \), there exists only one \( l \) such that \( M_k(m,l) = 1 \) and \( M_k(m,l') = 0 \) for all \( l' \neq l \).

We assume that for each fixed \( k \) and \( l \), \( \{M_k(m,l), m = 1, 2, \ldots\} \) is a sequence of i.i.d. random variables. This condition follows naturally from the usual assumption that \( \{t_{m+1} - t_m\} \) and \( \{\tau_{m,k}\}, k = 1, \ldots, K \) are sequences of i.i.d. random variables. Note that \( \sum_l M_k(m,l) = 1 \), which means that the delayed responses will be observed eventually, even though they may not be available immediately. Define \( \mu_{t,k} = E\{M_k(m,l)\} \). Thus, \( \sum_l \mu_{t,k} = 1 \) and \( \sum_{l=1}^{\infty} \mu_{t,k} \to 0 \) as \( l \to \infty \).

Regarding the convergence rate of \( \mu_{n,k} \), we make the following assumption.

**Assumption 1.** For some \( \varphi > 0 \),

\[
\sum_{l=t}^{\infty} \mu_{t,k} = P(\tau_{m,k} > t_{l+m} - t_m) = o(l^{-\varphi}), \quad k = 1, \ldots, K.
\]  

(3)

The probability in (3) is the probability that a patient on treatment \( k \) responds after at least another \( l \) patients have been assigned to treatments. Furthermore, it also provides a required
condition for the delay time $\tau_m(k)$. Assumption 1 implies that the delay time cannot be unreasonably large compared to the arrival time of the patients. The basic principle of this condition can also be found in a study of urn models with delayed responses of Bai, Hu & Rosenberger (2002), although only discrete responses that take finite possible values were considered in their study. A practical approach is to assume that the entry mechanism generates a Poisson process and the delay time has an exponential distribution in which both $\{\tau_m(k)\}$ and $\{t_{m+1} - t_m\}$ are sequences of i.i.d. exponential random variables with mean parameters, say, $\lambda_1 = \lambda_1(k) > 0$ and $\lambda_2 > 0$, respectively. This approach is popular in clinical studies, and the probability in (3) is simply $\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^l$.

3. ASYMPTOTIC RESULTS

Important asymptotic theorems are reported in this section, and the technical proofs are provided in the Appendix.

3.1. Some required conditions.

The following three groups of conditions are required for the theorems that are given in Section 3. The first group of conditions ((A1) and (A2)) is related to the moments of the response $\xi$. To obtain the consistency of $\mathcal{N}_n/n$, only the finiteness of the first moment (Condition (A1)) is needed. Condition (A2) is used to obtain the rate of the consistency and the asymptotic normality. The second group of conditions ((B1) and (B2)) deals with the allocation rule needed. Condition (A2) is used to obtain the rate of the consistency and the asymptotic normality. The second group of conditions ((B1) and (B2)) deals with the allocation rule $g(x, \rho)$. Here, we replace Conditions (B1)–(B3) in Hu & Zhang (2004a) by a new Condition (B1) that is easier to verify. Conditions (B1) and (B2) are satisfied if the allocation function is twice differentiable and the allocation functions are defined such that, if the current sample allocation proportion $N_{m,k}/m$ for a particular treatment $k$ exceeds the estimated target $\hat{\rho}_{m,k}$, then the next patient will be assigned to this treatment with a probability that is smaller than $\hat{\rho}_{m,k}$. For instance, the function that is defined in (7) and the allocation function that is proposed by Eisele & Woodroofe (1995) satisfy these conditions (see Hu & Zhang 2004a). The last group of conditions ((C1) and (C2)) involves the target allocation $\rho(\cdot)$. They are satisfied if $\rho(\cdot)$ is twice differentiable.

**CONDITION A.** Assume that the response sequence $\{\xi_n = (\xi_{n,1}, \ldots, \xi_{n,K})\}$ is a sequence of independent and identically distributed random vectors in $\mathbb{R}^K$ and that $\Theta = \mathbb{E}\xi_n$. Furthermore, assume that

(A1) $\mathbb{E}|\xi_{1,k}| < \infty$, $k = 1, \ldots, K$, and

(A2) for some $\varepsilon_0 > 0$, $\mathbb{E}|\xi_{1,k}|^{2+\varepsilon_0} < \infty$, $k = 1, \ldots, K$.

**CONDITION B.** The function of the allocation rule $g(x, y) = (g_1(x, y), \ldots, g_K(x, y))$ satisfies $g(v, v) = v$ and the following conditions:

(B1) There exists a constant $0 \leq \lambda_0 < 1$, such that for each $k = 1, \ldots, K$,

$$g_k(x, y) - y_k \leq \lambda_0 \quad \text{for all } x, y \text{ with } x_k > y_k, \quad x_1' = y_1' = 1.$$ 

(B2) There exists $\delta > 0$, for which the function $\rho(x, y)$ satisfies

$$g(x, y) = g(v, v) + \sum_{k=1}^K (x_k - v_k) \frac{\partial g}{\partial x_k} \bigg|_{(v, v)} + \sum_{k=1}^K (y_k - v_k) \frac{\partial g}{\partial y_k} \bigg|_{(v, v)} + o(\|x - v\|^{1+\delta}) + o(\|y - v\|^{1+\delta}) \quad \text{as } (x, y) \to (v, v).$$
CONDITION C. The proportion function \( z = (z_1, \ldots, z_d) \rightarrow \rho(z) : \mathbb{R}^K \rightarrow (0,1)^K \) with \( \rho(\Theta) = v \) and satisfies the following conditions:

(C1) \( \rho(z) \) is a continuous function.

(C2) There exists \( \delta > 0 \), for which

\[
\rho(z) = \rho(\Theta) + \sum_{k=1}^{K} (z_k - \theta_k) \frac{\partial \rho}{\partial z_k} \bigg|_{\Theta} + o(\|z - \Theta\|^{1+\delta}) \quad \text{as} \quad z \rightarrow \Theta.
\]

3.2. Consistency.

**Theorem 1** (Strong Consistency). Suppose that Conditions (A1), (B1), and (C1) are satisfied. If \( \sum_{i=1}^{\infty} \mu_{i,k} \rightarrow 0 \) as \( t \rightarrow \infty \), then \( N_n/n \rightarrow v \) a.s., and \( \Theta_n \rightarrow \Theta \) a.s. and \( \hat{\rho}_n \rightarrow v \) a.s.

The rates of consistency will depend on the derivatives of the allocation function \( g \). We write

\[
H = \left. \frac{\partial g}{\partial x} \right|_{(v,v)} = \left. \left( \frac{\partial g_j(x,y)}{\partial x_i} \right)_{i,j=1}^{K} \right|_{(v,v)}, \quad E = \left. \frac{\partial g}{\partial y} \right|_{(v,v)},
\]

where \( x = (x_1, \ldots, x_K) \) and \( y = (y_1, \ldots, y_K) \). It can easily be seen that \( H1' = E1' = 0' \) and \( \lambda_1 = 0 \) is an eigenvalue of both \( H \) and \( E \). We let \( \lambda_2, \ldots, \lambda_K \) be the other \( K-1 \) eigenvalues of \( H \), and let \( \lambda = \max\{\text{Re} (\lambda_2), \ldots, \text{Re} (\lambda_K)\} \).

**Theorem 2** (Rates of Consistency). Suppose that \( N_n/n \rightarrow v \) a.s., that Conditions (A2), (B2), and (C2) and Assumption 1 are satisfied, and that \( \lambda < 1 \). Let \( \varphi' = (2+\varepsilon_0)/(2+\varepsilon_0+(1+\varepsilon_0)\varphi) \). Then, for any \( \kappa > \max\{1/2, \lambda, \varphi'\} \),

\[
n^{-\kappa}(N_n - nv) \rightarrow 0 \quad \text{a.s. and} \quad \hat{\rho}_n - v = O\left(\sqrt{\frac{\log \log n}{n}}\right) + o(n^{\varphi'-1}(\log n)^2) \quad \text{a.s.}
\]

Furthermore, if \( \lambda < 1/2 \) and \( \varphi > (2+\varepsilon_0)/(1+\varepsilon_0) \), then

\[
N_n - nv = O\left(\sqrt{n \log \log n}\right) \quad \text{and} \quad \hat{\rho}_n - v = O\left(\sqrt{\frac{\log \log n}{n}}\right) \quad \text{a.s.} \quad (4)
\]

**Remark 1.** In Theorem 1, Assumption 1 about the response delay time has not been used, whereas in Theorem 2 it is required. This indicates that even though response delay does not affect the consistency property of the allocation proportions \( N_n/n \), it plays an important role in determining the rate of consistency.

3.3. Asymptotic normality.

To discuss normality, we need to introduce the following notation. Denote \( \sigma^2_k = \text{var}(\xi_{1,k}) \), \( k = 1, \ldots, K \), \( V = \text{diag}(\sigma^2_1/v_1, \ldots, \sigma^2_K/v_K) \), \( \Sigma_1 = \text{diag}(v) - v'v \),

\[
\Sigma_3 = \left( \frac{\partial \rho}{\partial \Theta} \right)' V \left( \frac{\partial \rho}{\partial \Theta} \right) = \sum_{k=1}^{K} \frac{1}{v_k} \left( \frac{\partial \rho}{\partial \theta_k} \right)' \sigma^2_k \left( \frac{\partial \rho}{\partial \theta_k} \right), \quad \Sigma_4 = E' \Sigma_3 E.
\]
THEOREM 3 (Asymptotic Normality). Suppose that $N_n/n \rightarrow v$ a.s. and Conditions (A2), (B2), and (C2) and Assumption 1 are satisfied. If $\lambda < 1/2$ and $\varphi > (2 + \varepsilon_0)/(1 + \varepsilon_0)$, then

$$n^{1/2}(N_n/n - v, \hat{\rho}_n - v) \xrightarrow{D} N(0, \Lambda),$$

(5)

where

$$\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Sigma_3 \end{pmatrix}$$

and

$$\Lambda_{11} = \int_0^1 \left( \frac{1}{x} \right)^H \Sigma_1 \left( \frac{1}{x} \right)^H dx + \int_0^1 dx \left[ \int_x^1 \left( \frac{1}{y} \right)^H dy \right] \Sigma_2 \left[ \int_x^1 \left( \frac{1}{y} \right)^H dy \right]$$

$$\Lambda_{12} = \Lambda_{21} = \Sigma_3 E(I - H)^{-1}.$$

We now apply Theorem 3 to the important class of DBCD treatment allocation functions that is given in Hu & Zhang (2004a). Their proposed treatment allocation function is

$$g_k(x, y) = \frac{y_k \left( \frac{y_k}{x_k} \right)^\gamma}{\sum_{j=1}^K y_j \left( \frac{y_j}{x_j} \right)^\gamma}, \quad k = 1, \ldots, K,$$

(7)

where $\gamma \geq 0$ is a constant that controls the degree of randomness of the procedure, from most random when $\gamma = 0$ to deterministic when $\gamma = \infty$.

COROLLARY 1. Let the allocation function $g(x, y)$ be defined as in (7). Suppose that Conditions (A2) and (C2) and Assumption 1 are satisfied, and that $\varphi > (2 + \varepsilon_0)/(1 + \varepsilon_0)$. Then, (4) and (5) hold with

$$\Lambda = \begin{pmatrix} \Sigma_1 & \Sigma_3 \\ \Sigma_3 & \Sigma_3 \end{pmatrix} \quad \text{and} \quad \Sigma = \frac{1}{1 + 2\gamma} \Sigma_1 + \frac{2(1 + \gamma)}{1 + 2\gamma} \Sigma_3.$$

(8)

The next corollary follows from the proof of the theorems.

COROLLARY 2. Suppose that $N_n/n \rightarrow v$ a.s. In addition, Conditions (A2), (B2), and (C2) and Assumption 1 are satisfied. If $\lambda < 1/2$ and $\varphi > (2 + \varepsilon_0)/(1 + \varepsilon_0)$, then for some $\varphi'_0 > 0$,

$$\hat{\Theta}_n - \tilde{\Theta}_n = o(n^{-1/2-\varphi'_0}) \quad \text{a.s.}$$

Remark 2. Corollary 2 ensures that the estimator $\hat{\Theta}_n$ with delayed responses has the same asymptotic properties (including strong consistency and asymptotic normality) as the estimator $\tilde{\Theta}_n$ (with the assumption of immediate responses) under the condition of a delayed rate. It is then obvious that the delayed rate assumption is satisfied when $\varphi \geq 2$.

The covariance matrix $\Lambda$ in (6) and (8) can be estimated by the plug-in method. For example, when the conditions in Corollary 1 are satisfied, the plug-in estimate of $\Lambda$ is defined using the following procedure. First, estimate $\sigma_k^2$ by the sample variance $\tilde{\sigma}_{n,k}^2$ of the observed responses on treatment $k$. Then, estimate $\Sigma_1$ and $\Sigma_3$, respectively, by
\[ \tilde{\Sigma}_{n,1} = \text{diag}(\rho(\tilde{\Theta}_n)) - (\rho(\tilde{\Theta}_n))'\rho(\tilde{\Theta}_n) \quad \text{and} \]
\[ \tilde{\Sigma}_{n,3} = \sum_{k=1}^{K} \frac{1}{\rho_k(\Theta_n)} \left( \frac{\partial \rho}{\partial \theta_k} \bigg|_{\Theta_n} \right)' \tilde{\sigma}_{n,k}^2 \left( \frac{\partial \rho}{\partial \theta_k} \bigg|_{\Theta_n} \right). \]

Finally, estimate \( \Lambda \) by
\[ \tilde{\Lambda}_n = \left( \frac{1}{1 + 2\gamma} \tilde{\Sigma}_{n,1} + \frac{2(1 + \gamma)}{1 + 2\gamma} \tilde{\Sigma}_{n,3} \right)^{-1} \tilde{\Sigma}_{n,3} \tilde{\sigma}_{n,3}^2. \]

The following theorem gives the consistency property of the proposed variance estimator.

**Theorem 4.** Under the conditions in Corollary 1,
\[ \tilde{\sigma}_{n,k}^2 \to \sigma_k^2 \quad \text{a.s.,} \quad k = 1, \ldots, K, \quad \text{and} \quad \tilde{\Lambda}_n \to \Lambda \quad \text{a.s.} \]

Further, if \( \varphi \geq 2 \) and \( \mathbb{E} |\xi_{1,k}|^{4+\varepsilon} < \infty \), \( k = 1, \ldots, K \), then
\[ \tilde{\sigma}_{n,k}^2 - \sigma_k^2 = O \left( \sqrt{\frac{\log \log n}{n}} \right) \quad \text{a.s.,} \quad k = 1, \ldots, K \quad \text{and} \quad \tilde{\Lambda}_n - \Lambda = O \left( \sqrt{\frac{\log \log n}{n}} \right) \quad \text{a.s.} \]

4. **Simulation Study**

In this section, we present an account of a simulation study that investigates the performance of the DBCD for clinical trials with delayed responses. For illustrative purposes, we simply take \( K = 2 \) and let the responses be dichotomous. We assume that the probability of success for treatment \( k \) is \( p_k = 1 - q_k \), \( k = 1, 2 \). Let \( \nu_k \) be the allocation target of treatment \( k \). Three different allocation targets are included in this study, as follows.

(a) Urn allocation proportion

The allocation target of \( N_{n1}/n \) is
\[ \nu_1 = \frac{q_2}{q_1 + q_2}. \]

By the formulas in Section 3, we have
\[ \text{var}(N_{n1}/n) \approx \frac{1}{n} \left( \frac{1}{1 + 2\gamma} \nu_1 (1 - \nu_1) + \frac{2(1 + \gamma)}{1 + 2\gamma} V_3 \right), \quad (9) \]
where
\[ V_3 = \frac{q_1 q_2 (p_1 + p_2)}{(q_1 + q_2)^3}. \]

(b) The RSIHR proportion of Rosenberger et al. (2001)

The allocation target of \( N_{n1}/n \) is
\[ \nu_1 = \frac{\sqrt{p_1}}{\sqrt{p_1} + \sqrt{p_2}}. \]
The variance of the allocation proportion is the same as that in Equation (9) with
\[ V_3 = \frac{1}{4\left(\sqrt{p_1} + \sqrt{p_2}\right)^3 \left(\frac{p_2q_1}{\sqrt{p_1}} + \frac{p_1q_2}{\sqrt{p_2}}\right)} \]

The RSIHR allocation of Rosenberger et al. (2001) is optimal in that it minimizes the expected number of treatment failures if the variance of the test (\(H_0: p_1 = p_2\)) statistic is fixed. However, for other criteria, it is no longer optimal.

(c) Neyman proportion

The allocation target of \(N_{n1}/n\) is
\[ \nu_1 = \frac{\sqrt{p_1q_1}}{\sqrt{p_1q_1} + \sqrt{p_2q_2}}. \]

The variance of the allocation proportion is the same as that in Equation (9) with
\[ V_3 = \frac{1}{4\left(\sqrt{p_1q_1} + \sqrt{p_2q_2}\right)^3 \left(\frac{p_2q_1(q_1 - p_1)^2}{\sqrt{p_1q_1}} + \frac{p_1q_1(q_2 - p_2)^2}{\sqrt{p_2q_2}}\right)} \]

The popular exponential distributions are used for both the delay times for the two treatments and the patient entry times. The mean parameters of the delay times for treatments 1 and 2 are \(\lambda_1\) and \(\lambda_2\), respectively, and the mean parameter for the patient entry times is \(\lambda_3\). There are four different configurations of the mean parameters. The first corresponds to a case in which there are no delayed responses, which means that all of the outcomes are observed with no time. The second corresponds to \((\lambda_1, \lambda_2, \lambda_3) = (1, 1, 1)\), which represents similar but moderate delay times for the responses of the two treatments. The third configuration is \((\lambda_1, \lambda_2, \lambda_3) = (5, 1, 1)\), which represents a larger difference in delay times for the responses of the two treatments. For example, one treatment is a surgical procedure, and the other one is the administration of a certain type of drug. The former usually yields relatively faster responses. Finally, we select \((\lambda_1, \lambda_2, \lambda_3) = (10, 10, 1)\), which stands for relatively large but identical delay times for both treatments.

In the simulation, we use \(\gamma = 2\) in the allocation function (7), as suggested by Rosenberger & Hu (2004). The number of subjects \(n\) is chosen to be 50 or 200, and the number of replications is 100,000. Let the average sample allocation proportions of the subjects that are being allocated to treatment 1 be \(p_1\), as tabulated in Tables 1 (\(n = 50\)) and 2 (\(n = 200\)). As the patients are allocated to either treatment 1 or treatment 2, the average sample allocation proportions of subjects who are being allocated to treatment 2 are not tabulated.

In addition to the allocation proportions, two standard deviations, \(SD_1\) and \(SD_2\), are also computed for each simulated allocation proportion. The standard deviation \(SD_1\) refers to the sample standard deviation that is computed based on the 100,000 simulated allocation proportions. For \(SD_2\), the plug-in method is employed. For each replication, the estimated values of \(p_1\) and \(p_2\) are obtained using sample success proportions. Then, Equation (9) is utilized to evaluate the standard deviation. After 100,000 simulation replications, the average standard deviation, \(SD_2\), is computed. The difference between \(SD_1\) and \(SD_2\) provides the basis for the evaluation of our variance estimates for various settings and sample sizes. If \(SD_1\) and \(SD_2\) are relatively close, then this will provide a simple approach to obtain a rough estimate of the standard deviation of the treatment allocation proportion in clinical trials.

The estimated allocation proportions are in general quite close to the target proportion. When \(n = 50\), the differences are in general larger, especially for urn allocation with very different success probabilities \((p_1, p_2)\). For immediate responses, the differences between the sample allocation proportions and the target allocation proportions are very small. This aligns with a number...
of recent research findings on the DBCD. Nevertheless, in the case of delayed responses, the differences become larger, especially for relatively large delay times \((\lambda_1, \lambda_2, \lambda_3) = (10, 10, 1)\).

### Table 1: Simulated allocation proportions with three different allocation targets \((n = 50)\).

<table>
<thead>
<tr>
<th>(p_1, p_2)</th>
<th>(\nu_1)</th>
<th>(\hat{\nu}_1) (SD(_1), SD(_2))</th>
<th>(\hat{\nu}_1) (SD(_1), SD(_2))</th>
<th>(\nu_1)</th>
<th>(\hat{\nu}_1) (SD(_1), SD(_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Immediate Response</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.8, .8</td>
<td>.500</td>
<td>.500 (.1317, .1415)</td>
<td>.500</td>
<td>.500 (.0375, .0381)</td>
<td>.500</td>
</tr>
<tr>
<td>.8, .6</td>
<td>.667</td>
<td>.641 (.1001, .1101)</td>
<td>.536</td>
<td>.535 (.0417, .0424)</td>
<td>.450</td>
</tr>
<tr>
<td>.8, .3</td>
<td>.778</td>
<td>.741 (.0678, .0779)</td>
<td>.620</td>
<td>.613 (.0520, .0575)</td>
<td>.466</td>
</tr>
<tr>
<td>.7, .5</td>
<td>.625</td>
<td>.608 (.0857, .0944)</td>
<td>.542</td>
<td>.541 (.0460, .0470)</td>
<td>.478</td>
</tr>
<tr>
<td>.7, .3</td>
<td>.700</td>
<td>.675 (.0687, .0774)</td>
<td>.604</td>
<td>.598 (.0534, .0585)</td>
<td>.500</td>
</tr>
<tr>
<td>.5, .5</td>
<td>.500</td>
<td>.500 (.0751, .0825)</td>
<td>.500</td>
<td>.500 (.0504, .0514)</td>
<td>.500</td>
</tr>
<tr>
<td>.5, .2</td>
<td>.615</td>
<td>.602 (.0581, .0644)</td>
<td>.613</td>
<td>.599 (.0600, .0720)</td>
<td>.556</td>
</tr>
<tr>
<td>.2, .2</td>
<td>.500</td>
<td>.500 (.0463, .0517)</td>
<td>.500</td>
<td>.500 (.0716, .0865)</td>
<td>.500</td>
</tr>
<tr>
<td>((\lambda_1, \lambda_2, \lambda_3) = (1, 1, 1))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.8, .8</td>
<td>.500</td>
<td>.500 (.1317, .1416)</td>
<td>.500</td>
<td>.500 (.0377, .0381)</td>
<td>.500</td>
</tr>
<tr>
<td>.8, .6</td>
<td>.667</td>
<td>.639 (.1007, .1101)</td>
<td>.536</td>
<td>.535 (.0418, .0424)</td>
<td>.450</td>
</tr>
<tr>
<td>.8, .3</td>
<td>.778</td>
<td>.741 (.0683, .0779)</td>
<td>.620</td>
<td>.613 (.0521, .0575)</td>
<td>.466</td>
</tr>
<tr>
<td>.7, .5</td>
<td>.625</td>
<td>.607 (.0861, .0944)</td>
<td>.542</td>
<td>.541 (.0462, .0470)</td>
<td>.478</td>
</tr>
<tr>
<td>.7, .3</td>
<td>.700</td>
<td>.674 (.0692, .0774)</td>
<td>.604</td>
<td>.598 (.0535, .0585)</td>
<td>.500</td>
</tr>
<tr>
<td>.5, .5</td>
<td>.500</td>
<td>.500 (.0754, .0825)</td>
<td>.500</td>
<td>.500 (.0506, .0514)</td>
<td>.500</td>
</tr>
<tr>
<td>.5, .2</td>
<td>.615</td>
<td>.602 (.0584, .0644)</td>
<td>.613</td>
<td>.599 (.0602, .0720)</td>
<td>.556</td>
</tr>
<tr>
<td>.2, .2</td>
<td>.500</td>
<td>.500 (.0466, .0517)</td>
<td>.500</td>
<td>.500 (.0717, .0865)</td>
<td>.500</td>
</tr>
<tr>
<td>((\lambda_1, \lambda_2, \lambda_3) = (5, 1, 1))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.8, .8</td>
<td>.500</td>
<td>.496 (.1323, .1417)</td>
<td>.500</td>
<td>.499 (.0379, .0381)</td>
<td>.500</td>
</tr>
<tr>
<td>.8, .6</td>
<td>.667</td>
<td>.639 (.1017, .1101)</td>
<td>.536</td>
<td>.534 (.0420, .0424)</td>
<td>.450</td>
</tr>
<tr>
<td>.8, .3</td>
<td>.778</td>
<td>.741 (.0697, .0780)</td>
<td>.620</td>
<td>.612 (.0523, .0575)</td>
<td>.466</td>
</tr>
<tr>
<td>.7, .5</td>
<td>.625</td>
<td>.605 (.0873, .0944)</td>
<td>.542</td>
<td>.540 (.0465, .0470)</td>
<td>.478</td>
</tr>
<tr>
<td>.7, .3</td>
<td>.700</td>
<td>.673 (.0708, .0774)</td>
<td>.604</td>
<td>.597 (.0538, .0585)</td>
<td>.500</td>
</tr>
<tr>
<td>.5, .5</td>
<td>.500</td>
<td>.501 (.0765, .0825)</td>
<td>.500</td>
<td>.500 (.0512, .0514)</td>
<td>.500</td>
</tr>
<tr>
<td>.5, .2</td>
<td>.615</td>
<td>.602 (.0596, .0644)</td>
<td>.613</td>
<td>.598 (.0607, .0720)</td>
<td>.556</td>
</tr>
<tr>
<td>.2, .2</td>
<td>.500</td>
<td>.502 (.0472, .0518)</td>
<td>.500</td>
<td>.502 (.0718, .0864)</td>
<td>.500</td>
</tr>
<tr>
<td>((\lambda_1, \lambda_2, \lambda_3) = (10, 10, 1))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.8, .8</td>
<td>.500</td>
<td>.500 (.1341, .1421)</td>
<td>.500</td>
<td>.500 (.0390, .0381)</td>
<td>.500</td>
</tr>
<tr>
<td>.8, .6</td>
<td>.667</td>
<td>.629 (.1054, .1104)</td>
<td>.536</td>
<td>.534 (.0437, .0424)</td>
<td>.450</td>
</tr>
<tr>
<td>.8, .3</td>
<td>.778</td>
<td>.727 (.0735, .0780)</td>
<td>.620</td>
<td>.608 (.0533, .0574)</td>
<td>.466</td>
</tr>
<tr>
<td>.7, .5</td>
<td>.625</td>
<td>.601 (.0915, .0945)</td>
<td>.542</td>
<td>.539 (.0486, .0470)</td>
<td>.478</td>
</tr>
<tr>
<td>.7, .3</td>
<td>.700</td>
<td>.666 (.0746, .0775)</td>
<td>.604</td>
<td>.593 (.0551, .0584)</td>
<td>.500</td>
</tr>
<tr>
<td>.5, .5</td>
<td>.500</td>
<td>.500 (.0805, .0826)</td>
<td>.500</td>
<td>.500 (.0533, .0514)</td>
<td>.500</td>
</tr>
<tr>
<td>.5, .2</td>
<td>.615</td>
<td>.598 (.0624, .0644)</td>
<td>.613</td>
<td>.592 (.0614, .0729)</td>
<td>.556</td>
</tr>
<tr>
<td>.2, .2</td>
<td>.500</td>
<td>.500 (.0493, .0517)</td>
<td>.500</td>
<td>.500 (.0721, .0862)</td>
<td>.500</td>
</tr>
</tbody>
</table>

SD\(_1\) is the simulated standard deviation, and SD\(_2\) is the estimated standard deviation via the formulas in Section 3.
### Table 2: Simulated allocation proportions with three different allocation targets ($n = 200$).

<table>
<thead>
<tr>
<th>$p_1, p_2$</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Immediate Response</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.8, .8</td>
<td>.500</td>
<td>(.0758, .0766)</td>
<td>.500</td>
<td>(.0186, .0187)</td>
<td>.500</td>
<td>(.0337, .0336)</td>
</tr>
<tr>
<td>.8, .6</td>
<td>.667</td>
<td>(.0557, .0570)</td>
<td>.536</td>
<td>(.0207, .0208)</td>
<td>.450</td>
<td>(.0276, .0275)</td>
</tr>
<tr>
<td>.8, .3</td>
<td>.778</td>
<td>(.0369, .0382)</td>
<td>.620</td>
<td>(.0280, .0282)</td>
<td>.466</td>
<td>(.0294, .0292)</td>
</tr>
<tr>
<td>.7, .5</td>
<td>.625</td>
<td>(.0468, .0481)</td>
<td>.542</td>
<td>(.0229, .0230)</td>
<td>.478</td>
<td>(.0204, .0203)</td>
</tr>
<tr>
<td>.7, .3</td>
<td>.700</td>
<td>(.0373, .0384)</td>
<td>.604</td>
<td>(.0285, .0286)</td>
<td>.500</td>
<td>(.0236, .0235)</td>
</tr>
<tr>
<td>.5, .5</td>
<td>.500</td>
<td>(.0405, .0417)</td>
<td>.500</td>
<td>(.0250, .0252)</td>
<td>.500</td>
<td>(.0162, .0163)</td>
</tr>
<tr>
<td>.5, .2</td>
<td>.615</td>
<td>(.0309, .0318)</td>
<td>.613</td>
<td>(.0353, .0357)</td>
<td>.556</td>
<td>(.0271, .0270)</td>
</tr>
<tr>
<td>.2, .2</td>
<td>.500</td>
<td>(.0245, .0252)</td>
<td>.500</td>
<td>(.0419, .0423)</td>
<td>.500</td>
<td>(.0335, .0336)</td>
</tr>
</tbody>
</table>

$\lambda_1$, $\lambda_2$, $\lambda_3 = (1, 1, 1)$

<table>
<thead>
<tr>
<th>$p_1, p_2$</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>.8, .8</td>
<td>.500</td>
<td>(.0761, .0767)</td>
<td>.500</td>
<td>(.0186, .0187)</td>
<td>.500</td>
<td>(.0338, .0336)</td>
</tr>
<tr>
<td>.8, .6</td>
<td>.667</td>
<td>(.0558, .0570)</td>
<td>.536</td>
<td>(.0207, .0208)</td>
<td>.450</td>
<td>(.0276, .0275)</td>
</tr>
<tr>
<td>.8, .3</td>
<td>.778</td>
<td>(.0370, .0382)</td>
<td>.620</td>
<td>(.0281, .0282)</td>
<td>.466</td>
<td>(.0294, .0292)</td>
</tr>
<tr>
<td>.7, .5</td>
<td>.625</td>
<td>(.0469, .0481)</td>
<td>.542</td>
<td>(.0229, .0230)</td>
<td>.478</td>
<td>(.0204, .0203)</td>
</tr>
<tr>
<td>.7, .3</td>
<td>.700</td>
<td>(.0374, .0384)</td>
<td>.604</td>
<td>(.0285, .0286)</td>
<td>.500</td>
<td>(.0236, .0235)</td>
</tr>
<tr>
<td>.5, .5</td>
<td>.500</td>
<td>(.0406, .0417)</td>
<td>.500</td>
<td>(.0251, .0252)</td>
<td>.500</td>
<td>(.0162, .0163)</td>
</tr>
<tr>
<td>.5, .2</td>
<td>.615</td>
<td>(.0309, .0318)</td>
<td>.613</td>
<td>(.0354, .0357)</td>
<td>.556</td>
<td>(.0272, .0270)</td>
</tr>
<tr>
<td>.2, .2</td>
<td>.500</td>
<td>(.0245, .0250)</td>
<td>.500</td>
<td>(.0420, .0423)</td>
<td>.500</td>
<td>(.0336, .0336)</td>
</tr>
</tbody>
</table>

$\lambda_1$, $\lambda_2$, $\lambda_3 = (5, 1, 1)$

<table>
<thead>
<tr>
<th>$p_1, p_2$</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>.8, .8</td>
<td>.500</td>
<td>(.0766, .0767)</td>
<td>.500</td>
<td>(.0186, .0187)</td>
<td>.500</td>
<td>(.0339, .0336)</td>
</tr>
<tr>
<td>.8, .6</td>
<td>.667</td>
<td>(.0562, .0570)</td>
<td>.536</td>
<td>(.0207, .0208)</td>
<td>.450</td>
<td>(.0279, .0275)</td>
</tr>
<tr>
<td>.8, .3</td>
<td>.778</td>
<td>(.0373, .0382)</td>
<td>.620</td>
<td>(.0281, .0282)</td>
<td>.466</td>
<td>(.0296, .0292)</td>
</tr>
<tr>
<td>.7, .5</td>
<td>.625</td>
<td>(.0472, .0481)</td>
<td>.542</td>
<td>(.0230, .0230)</td>
<td>.478</td>
<td>(.0205, .0203)</td>
</tr>
<tr>
<td>.7, .3</td>
<td>.700</td>
<td>(.0376, .0384)</td>
<td>.604</td>
<td>(.0286, .0286)</td>
<td>.500</td>
<td>(.0237, .0235)</td>
</tr>
<tr>
<td>.5, .5</td>
<td>.500</td>
<td>(.0408, .0417)</td>
<td>.500</td>
<td>(.0252, .0252)</td>
<td>.500</td>
<td>(.0162, .0163)</td>
</tr>
<tr>
<td>.5, .2</td>
<td>.615</td>
<td>(.0311, .0318)</td>
<td>.613</td>
<td>(.0355, .0357)</td>
<td>.556</td>
<td>(.0272, .0270)</td>
</tr>
<tr>
<td>.2, .2</td>
<td>.500</td>
<td>(.0246, .0250)</td>
<td>.500</td>
<td>(.0422, .0423)</td>
<td>.500</td>
<td>(.0337, .0336)</td>
</tr>
</tbody>
</table>

$\lambda_1$, $\lambda_2$, $\lambda_3 = (10, 10, 1)$

<table>
<thead>
<tr>
<th>$p_1, p_2$</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
<th>$\nu_1$</th>
<th>$\hat{\nu}_1$ (SD1, SD2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>.8, .8</td>
<td>.500</td>
<td>(.0779, .0767)</td>
<td>.500</td>
<td>(.0187, .0187)</td>
<td>.500</td>
<td>(.0346, .0336)</td>
</tr>
<tr>
<td>.8, .6</td>
<td>.667</td>
<td>(.0571, .0570)</td>
<td>.536</td>
<td>(.0209, .0208)</td>
<td>.450</td>
<td>(.0282, .0275)</td>
</tr>
<tr>
<td>.8, .3</td>
<td>.778</td>
<td>(.0378, .0383)</td>
<td>.620</td>
<td>(.0286, .0282)</td>
<td>.466</td>
<td>(.0300, .0292)</td>
</tr>
<tr>
<td>.7, .5</td>
<td>.625</td>
<td>(.0480, .0481)</td>
<td>.542</td>
<td>(.0233, .0230)</td>
<td>.478</td>
<td>(.0207, .0203)</td>
</tr>
<tr>
<td>.7, .3</td>
<td>.700</td>
<td>(.0382, .0384)</td>
<td>.604</td>
<td>(.0291, .0286)</td>
<td>.500</td>
<td>(.0240, .0235)</td>
</tr>
<tr>
<td>.5, .5</td>
<td>.500</td>
<td>(.0415, .0417)</td>
<td>.500</td>
<td>(.0255, .0252)</td>
<td>.500</td>
<td>(.0162, .0163)</td>
</tr>
<tr>
<td>.5, .2</td>
<td>.615</td>
<td>(.0315, .0318)</td>
<td>.613</td>
<td>(.0360, .0357)</td>
<td>.556</td>
<td>(.0276, .0270)</td>
</tr>
<tr>
<td>.2, .2</td>
<td>.500</td>
<td>(.0249, .0252)</td>
<td>.500</td>
<td>(.0430, .0423)</td>
<td>.500</td>
<td>(.0342, .0336)</td>
</tr>
</tbody>
</table>

$SD_1$ is the simulated standard deviation, and $SD_2$ is the estimated standard deviation via the formulas in Section 3.

As for the variance estimates, our proposed method works quite well for $n = 200$ when $SD_2$ is very close to $SD_1$ and reasonably well for $n = 50$, regardless of the degree of delay.
time. This indicates that the proposed estimation procedure gives a relatively good approximation of the actual variances, and, more importantly, shows that delayed responses do not affect the integrity of the variance estimates. This is the crucial advantage of the DBCD in the generation of treatment allocation schemes.

To provide a better understanding of the effect of the degree of delay time on the allocation proportions, another simulation study was conducted for \( p_1 = 0.8, p_2 = 0.3, n = 50 \) for the urn proportion. The disparities between \( \nu \) and \( \hat{\nu} \) are larger for the urn allocation, whereas the differences for the other two allocation schemes are smaller. Hence, the corresponding expected simulation results for the allocation of Rosenberger et al. (2001) and the Neyman allocation are not reported here. The delay parameters are \( (\lambda_1, \lambda_2, \lambda_3) = (\lambda^*, 0, 1) \). Therefore, we assume an immediate response for the second treatment. For treatment 1, the delay time is \( \lambda^* = 0, 1, \ldots, 30 \). The estimated bias is taken to be \( \nu_1 - \hat{\nu}_1 \). As shown in Figure 1, SD1 varies slightly as \( \lambda^* \) increases. As for the estimated bias, it increases gradually as \( \lambda^* \) increases. However, unless \( \lambda^* \) is very large, the bias is not substantial. Again, the findings support the idea that doubly adaptive biased coin designs are relatively insensitive to delay time.

![Figure 1: Estimated bias and standard deviation (SD1) for different delay times \( \lambda^* \) [Urn proportion with \( p_1 = 0.8, p_2 = 0.3, n = 50 \).](image)

5. CONCLUSION

This paper presents two important theoretical results. First, under widely satisfied conditions and a delay mechanism, we show that the asymptotic properties of doubly adaptive biased coin designs remain insensitive to a stochastic delay in updating the sequential estimator of the unknown parameters. Second, in terms of the estimation procedure, we can employ standard maximum likelihood estimation, moment estimation, or other consistent statistical estimators to perform the necessary inferences even when responses are available with some delay, as long as certain standard clinical trial conditions are satisfied. This is a direct result of Theorem 1 and Corollary 2 of this paper and of Lemma 1 of Hu, Rosenberger & Zhang (2006).

Bai, Hu & Rosenberger (2002) and Hu & Zhang (2004b) considered the asymptotic properties of the generalized Friedman’s urn with multinomial responses. However, they did not
examine the properties of estimators with a delay model. The results in this paper, in contrast, can be applied to both discrete and continuous responses. Corollary 2 shows that the asymptotic properties of the estimator $\tilde{\Theta}_n$ are not affected by delayed responses under widely satisfied conditions for the delay mechanism. The theoretical results in this paper are based on the assumption that the sample size is large, but, in practice, the delay mechanism may be influential in a fixed sample size. As demonstrated by our simulation study, the delay mechanism plays a larger role for a small sample size ($n = 50$) than for a moderate sample size ($n = 200$).

APPENDIX

Proofs of the asymptotic theorems.

Theorems 4.1–4.3 of Hu & Zhang (2004a) proved asymptotic theories that are similar to our theorems for designs without delayed responses. Hence, the proofs of Theorems 1–3 in this paper can be completed by showing that the two estimates $\tilde{\Theta}_n$ and $\tilde{\Theta}_n$ (the estimate without delayed responses) are equivalent under the given conditions. To accomplish this task, we need three additional lemmas.

First, note that from the time patient $j$ is assigned a treatment to the time patient $(j + 1)$ arrives, all of the responses of the patients on treatment $k$ that we have observed correspond to the nonzero $X_{j - m, k}M_k(j - m, m)\xi_{j - m, k}, m = 0, \ldots, j - 1$, and the total number of such cases is

$$t_{j, k} := \sum_{m=0}^{j-1} X_{j-m,k}M_k(j - m, m) = \sum_{m=1}^{j} M_k(m, j - m)X_{m,k}.$$ 

Set

$$s_{j,k} = \sum_{m=0}^{j-1} X_{j-m,k}M_k(j - m, m)\xi_{j-m,k} = \sum_{m=1}^{j} M_k(m, j - m)X_{m,k}\xi_{m,k}.$$ 

Remember that $S_{n,k}$ is the total sum of the responses on treatment $k$ that are observed up to the time patient $(n + 1)$ is ready for treatment allocation, and $T_{n,k}$ is the number of responses observed. Then, for each $k = 1, \ldots, K$,

$$S_{n,k} = \sum_{j=1}^{n} s_{j,k} = \sum_{j=1}^{n} \sum_{m=1}^{j} M_k(m, j - m)X_{m,k}\xi_{m,k}$$

$$= \sum_{m=1}^{n} \sum_{j=m}^{n} M_k(m, j - m)X_{m,k}\xi_{m,k} = \sum_{m=1}^{n} \sum_{j=0}^{n-m} X_{m,k}M_k(m, j)\xi_{m,k}, \quad (10)$$

and similarly,

$$T_{n,k} = \sum_{m=1}^{n} t_{j,k} = \sum_{m=1}^{n} \sum_{j=0}^{n-m} M_k(m, j)X_{m,k}. \quad (11)$$

**Lemma 1.** Let $\{\eta_k, k = 1, 2, \ldots\}$ be a sequence of independent and identically distributed random variables with zero means and let $\{\varepsilon_k, k = 1, 2, \ldots\}$ be another sequence of random variables that takes only two values, 0 or 1. Denote $\zeta_n = \sum_{j=1}^{n} \varepsilon_j$. Suppose that for each $n$, $\eta_n$ is dependent on $\{\eta_1, \ldots, \eta_{n-1}, \varepsilon_1, \ldots, \varepsilon_n\}$. Then, in the event that $\{\zeta_n \to \infty\}$,

$$\frac{\sum_{j=1}^{n} \varepsilon_j \eta_j}{\zeta_n} \to 0 \quad a.s.$$

and

$$\limsup_{n \to \infty} \frac{|\sum_{j=1}^{n} \varepsilon_j \eta_j|}{\sqrt{2\zeta_n \log \log \zeta_n}} = (E \eta_1^2)^{1/2}, \quad \text{if } E \eta_1^2 < \infty.$$
Proof. The proof is similar to the proof of Lemma A.4 in Hu & Zhang (2004a).

**Lemma 2.** For the doubly adaptive biased coin design with delayed responses, if Conditions (A1) and (C1) are satisfied, then the event \( \{N_{n,k} \to \infty\} \) implies almost surely that \( T_{n,k}/N_{n,k} \to 1, \hat{\theta}_{n,k} \to \theta_k, \) and \( \theta_{n,k} \to \theta_k. \) Further, if Condition (A2) is also satisfied, then the event \( \{N_{n,k} \to \infty\} \) implies almost surely that \( \hat{\theta}_{n,k} - \theta_k = O\left(\sqrt{(\log \log N_{n,k})/N_{n,k}}\right). \)

Proof. It is obvious that

\[
\sum_{j=0}^{\infty} M_k(m, j) = 1 \quad \text{and} \quad T_{n,k} \leq N_{n,k} \quad \text{for } k = 1, \ldots, K. \tag{12}
\]

For each \( j \) and \( k, \) \( \{M_k(m, j), m = 1, 2, \ldots\} \) is a sequence of i.i.d. random variables. From Lemma 1, it follows that

\[
\{N_{n,k} \to \infty\} \implies \frac{\sum_{m=1}^{n} X_{m,k} M_k(m, j)}{N_{n,k}} \to E[M_k(1, j)] \quad \text{a.s., } j = 0, 1, \ldots.
\]

Then, in the event that \( \{N_{n,k} \to \infty\},
\[
\liminf_{n \to \infty} \frac{\sum_{m=1}^{n} \sum_{j=0}^{n-m} X_{m,k} M_k(m, j)}{N_{n,k}} \geq \frac{\sum_{m=1}^{n} \sum_{j=0}^{r} X_{m,k} M_k(m, j)}{N_{n,k}}
\]

\[
= \sum_{j=0}^{r} E[M_k(1, j)] = E\left[\left(\sum_{j=0}^{r} M_k(1, j)\right)\right] \to 1 \quad \text{as } r \to \infty.
\]

We therefore conclude that in the event that \( \{N_{n,k} \to \infty\},
\[
\frac{T_{n,k}}{N_{n,k}} = \frac{\sum_{m=1}^{n} \sum_{j=0}^{n-m} X_{m,k} M_k(m, j)}{N_{n,k}} \to 1 \quad \text{a.s.} \tag{13}
\]

Note that
\[
\sum_{m=1}^{n} \sum_{j=n-m+1}^{\infty} X_{m,k} M_k(m, j) |\xi_{m,k}|
\]

\[
\leq A \sum_{m=1}^{n} \sum_{j=n-m+1}^{\infty} X_{m,k} M_k(m, j) + \sum_{m=1}^{n} \sum_{j=n-m+1}^{\infty} X_{m,k} M_k(m, j) |\xi_{m,k}| I\{|\xi_{m,k}| \geq A\}
\]

\[
\leq A \sum_{m=1}^{n} (N_{n,k} - T_{n,k}) + \sum_{m=1}^{n} X_{m,k} |\xi_{m,k}| I\{|\xi_{m,k}| \geq A\}.
\]

By Lemma 1 and (13), in the event that \( \{N_{n,k} \to \infty\},
\[
\limsup_{n \to \infty} \frac{\sum_{m=1}^{n} \sum_{j=n-m+1}^{\infty} X_{m,k} M_k(m, j) |\xi_{m,k}|}{N_{n,k}}
\]

\[
\leq 0 + E[|\xi_{1,k}| I\{|\xi_{1,k}| \geq A\}] \to 0 \quad \text{a.s., as } A \to \infty.
\]

Also, again by Lemma 1,
\[
\hat{\theta}_{n,k} = \frac{\sum_{m=1}^{n} X_{m,k} \xi_{m,k} + \theta_{0,k}}{N_{n,k} + 1} \to \theta_k \quad \text{a.s. in the event that } \{N_{n,k} \to \infty\}.
\]
Thus, by noting (10) and (12), we conclude that in the event that \( \{N_{n,k} \to \infty \} \),

\[
\frac{S_{n,k}}{N_{n,k}} = \frac{\sum_{m=1}^{n} X_{m,k} \xi_{m,k}}{N_{n,k}} - \sum_{m=1}^{n} \sum_{j=n-m+1}^{\infty} X_{m,k} M_k(m,j) \xi_{m,k} \to \theta_k \quad \text{a.s.} \tag{14}
\]

Combining (13), (14), and the definition of \( \hat{\theta}_{n,k} \) completes the proof of the lemma. \( \square \)

**Lemma 3.** For the doubly adaptive biased coin design with delayed responses, suppose that \( N_n/n \to v \) a.s. If Condition (A2) and Assumption 1 are satisfied, then for \( \varphi' = (2 + \varepsilon_0)/(2 + \varepsilon_0 + (1 + \varepsilon_0) \varepsilon_0) \),

\[
\tilde{\Theta}_n - \hat{\Theta}_n = o(n^{\varphi'-1}(\log n)^2) + o(n^{-1+\varepsilon_0+\varphi''}) \quad \text{a.s.}
\]

In particular, if \( \varphi > (2 + \varepsilon_0)/(1 + \varepsilon_0) \), then for some \( \varphi'' > 0 \),

\[
\tilde{\Theta}_n - \hat{\Theta}_n = o(n^{-1/2-\varphi''}) \quad \text{a.s.}
\]

**Proof.** Note that

\[
n/N_{n,k} \to 1/v_k < \infty \quad \text{a.s., } k = 1, \ldots, K.
\]

By (10) and (12), for each \( k = 1, \ldots, K \),

\[
S_{n,k} = \sum_{m=1}^{n} X_{m,k} \xi_{m,k} - \sum_{m=1}^{n} \sum_{j=n-m+1}^{\infty} X_{m,k} M_k(m,j) \xi_{m,k} := J_{n,k} - R_{n,k}.
\]

Let \( 0 < \varphi' < 1 \) be a number, the value of which will be defined later. Let \( l_n = \lceil n^{\varphi'} \rceil \) and \( I_k(m,p) = \sum_{j=p+1}^{\infty} M_k(m,j) \). Then,

\[
|R_{n,k}| = \left| \sum_{m=n-l_n+1}^{n} \sum_{j=n-m+1}^{\infty} M_k(m,j) X_{m,k} \xi_{m,k} + \sum_{m=1}^{n-l_n} \sum_{j=n-m+1}^{\infty} M_k(m,j) X_{m,k} \xi_{m,k} \right|
\]

\[
\leq \sum_{m=n-l_n+1}^{n} |\xi_{m,k}| + \sum_{m=1}^{n-l_n} \sum_{j=n-m+1}^{\infty} M_k(m,j) |\xi_{m,k}|
\]

\[
\leq \sum_{m=n-l_n+1}^{n} |\xi_{m,k}| + \sum_{m=1}^{n} I_k(m,l_m) |\xi_{m,k}|
\]

\[
\leq Cn^{\varphi'} + \sum_{m=n-l_n+1}^{n} (|\xi_{m,k}| - E(|\xi_{m,k}|)) + \sum_{m=1}^{n} I_k(m,l_m) |\xi_{m,k}|. \tag{17}
\]

For the second term, by Theorems 1.2.1 and 2.6.6 of Csörgő & Révész (1981),

\[
\sum_{m=n-l_n+1}^{n} (|\xi_{m,k}| - E(|\xi_{m,k}|)) = o(n^{\varphi'}) + o(n^{1/2+\varphi_0}) \quad \text{a.s.}
\]

For the third term in (17), by Assumption 1 and the Hölder inequality, we have

\[
\sum_{m=1}^{n} \frac{E[I_k(m,l_m)|\xi_{m,k}|]}{m^{\phi'(1+\varepsilon_0)}} (log m)^2
\]

\[
\leq \sum_{m=1}^{n} m^{-1} \frac{\phi'(1+\varepsilon_0)}{2+\varepsilon_0} (log m)^{-2} (E[I_k(m,l_m)])^{1+\varepsilon_0} (E|\xi_{m,k}|^{2+\varepsilon_0}) \frac{1}{2+\varepsilon_0}
\]

\[
\leq C \sum_{m=1}^{n} m^{-1} (log m)^{-2} < \infty.
\]
It follows that
\[ \sum_{m=1}^{n} I_k(m, l_m) |\xi_{m,k}| = o\left(n^{1-\frac{\varphi'(1+\delta_0)}{2+\delta_0}} (\log n)^2 \right) \quad \text{a.s.} \]
Combining these arguments, we have
\[ R_{n,k} = O(n^{\varphi'}) + o\left(n^{1-\frac{\varphi'(1+\delta_0)}{2+\delta_0}} (\log n)^2 \right) + o\left(n^{\frac{1+\delta_0}{2+\delta_0}} \right) \quad \text{a.s.} \]
Choosing
\[ \varphi' = \frac{2 + \varepsilon_0}{2 + \varepsilon_0 + (1 + \varepsilon_0)\varphi} \]
yields
\[ R_{n,k} = o\left(n^{\varphi'} (\log n)^2 \right) + o\left(n^{\frac{1+\delta_0}{2+\delta_0}} \right) \quad \text{a.s.} \tag{18} \]
Combining (15)–(18), we have
\[ \frac{S_{n,k}}{N_{n,k}} = \frac{J_{n,k}}{N_{n,k}} - \frac{R_{n,k}}{N_{n,k}} = \frac{J_{n,k}}{N_{n,k}} + o\left(n^{\varphi'-1} (\log n)^2 \right) + o\left(n^{\frac{1+\delta_0}{2+\delta_0}} \right) \quad \text{a.s.} \]
Similarly,
\[ \frac{T_{n,k}}{N_{n,k}} = 1 + o\left(n^{\varphi'-1} (\log n)^2 \right) + o\left(n^{\frac{1+\delta_0}{2+\delta_0}} \right) \quad \text{a.s.} \]
Consequently, for each \( k = 1, \ldots, K \),
\[ \hat{\theta}_{n,k} = \frac{S_{n,k} + \theta_{0,k}}{T_{n,k} + 1} = \frac{J_{n,k} + \theta_{0,k}}{N_{n,k} + 1} + o\left(n^{\varphi'-1} (\log n)^2 \right) + o\left(n^{\frac{1+\delta_0}{2+\delta_0}} \right) \]
\[ = \hat{\theta}_{n,k} + o\left(n^{\varphi'-1} (\log n)^2 \right) + o\left(n^{\frac{1+\delta_0}{2+\delta_0}} \right) \quad \text{a.s.} \]
The proof of Lemma 3 is complete. \( \square \)

**Proof of Theorem 1.** By Lemma 2, it follows that
\[ \hat{\theta}_{m,k} \rightarrow \theta_k \quad \text{a.s. in the event that } \{N_{m,k} \rightarrow \infty\}. \tag{19} \]
However, \( \hat{\theta}_{m,k} \) will fix to a value eventually in the event that \( \{\sup_m N_{m,k} < \infty\} \). In either case, \( \hat{\theta}_{m,k} \) has a limit \( \hat{\theta}^*_k \) in the parameter space, \( k = 1, \ldots, K \). By the continuity of \( \hat{\rho}(\cdot), \rho_{m-1} \rightarrow \rho(\theta_1^*, \ldots, \theta_K^*) := v^* \in (0, 1)^K \) a.s. Denote \( M_{n,k} = \sum_{m=1}^{n}(X_{m,k} - E[X_{m,k} | F_{m-1}]) \).
Then, \( \max_{m \leq n} |M_{m,k}|/n \rightarrow 0 \) a.s. by the law of large numbers of martingales. Let \( \nu_n = \max\{m \geq Km_0 + 1 : N_{m-1,k}/(m-1) \leq \hat{\rho}_{m-1,k} \} \) and \( \max\{0\} = Km_0 \). Note the allocation probability \( p_{m,k} \leq \hat{\rho}_{m-1,k} + \lambda_0(N_{m-1,k}/(m-1) - \hat{\rho}_{m-1,k}) \) if \( N_{m-1,k}/(m-1) > \hat{\rho}_{m-1,k} \) by Condition (B1). It follows that
\[ N_{n,k} = N_{\nu_n,k} = \sum_{m=\nu_n+1}^{n} (X_{m,k} - E[X_{m,k} | F_{m-1}]) + \sum_{m=\nu_n+1}^{n} p_{m,k} \]
\[ \leq 1 + N_{\nu_n-1,k} + M_{\nu_n,k} - M_{\nu_n,k} + \sum_{m=\nu_n}^{n-1} (\hat{\rho}_{m,k} + \lambda_0(N_{m,k}/m - \hat{\rho}_{m,k})) \]
\[ \leq 1 + N_{Km_0-1,k} + 2 \max_{m \leq n} |M_{m,k}| \]
\[ + (\nu_n - 1)\hat{\rho}_{\nu_n-1,k} + \sum_{m=\nu_n}^{n-1} \hat{\rho}_{m,k} + \lambda_0 \sum_{m=\nu_n}^{n-1} (N_{m,k}/m - \hat{\rho}_{m,k}). \]
It then follows that \( \limsup_{n \to \infty} N_{n,k}/n \leq v^*_k + \lambda_0 (\limsup_{n \to \infty} N_{n,k}/n - v^*_k) \) a.s. Note \( \lambda_0 < 1 \). We conclude that \( \limsup_{n \to \infty} N_{n,k}/n \leq v^*_k \) a.s., \( k = 1, \ldots, K \). From the fact that \( \sum_{k=1}^n N_{n,k}/n = \sum_{k=1}^K v^*_k = 1 \), we conclude that \( N_{n,k}/n \to v^*_k \) a.s., which implies that \( N_{n,k} \to \infty \) a.s., \( k = 1, \ldots, K \). Hence, the limits \( \theta^*_k \) and \( \theta_k \) (\( v^* \) and \( v \)) must be identical by (19). The proof is now complete. \( \square \)

Proofs of Theorems 2 and 3. By using Lemmas 2 and 3, the proofs follow closely to those given in Theorems 4.2 and 4.3 of Hu & Zhang (2004a). We therefore omit the details here.

Proof of Theorem 4. Observe that

\[
\hat{\Theta}_n - \Theta = O \left( \sqrt{\frac{\log \log n}{n}} \right) \quad \text{a.s.}
\]

due to Lemma 2, Lemma 3 and Equation (15), and also that

\[
\hat{\sigma}^2_{n,k} = \frac{1}{T_{n,k}} \sum_{m=1}^n \sum_{j=0}^{n-m} X_{m,k} M(m,j) (\xi_{m,k} - \hat{\theta}_{n,k})^2.
\]

It is sufficient to show that for \( k = 1, \ldots, K \),

\[
\tilde{V}^*_n = \frac{1}{T_{n,k}} \sum_{j=1}^{T_{n,k}} (\xi_{j,k} - \theta_k)^2
\]

\[
= \frac{1}{T_{n,k}} \sum_{m=1}^n \sum_{j=0}^{n-m} X_{m,k} M(m,j) (\xi_{m,k} - \theta_k)^2 \to \sigma^2_k \quad \text{a.s.,}
\]

and, on the condition that \( \mathbb{E} |\xi_{1,k}|^{4+\epsilon} < \infty \),

\[
\tilde{V}^*_n - \sigma^2_k = O \left( \sqrt{\frac{\log \log n}{n}} \right)
\]
ACKNOWLEDGEMENTS

Special thanks go to the anonymous referees, the associate editor and the editor for their constructive comments, which led to a much improved version of the paper. This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project no CUHK400608). F. Hu’s work was supported in part by National Science Foundation Grant (U.S.A.) DMS-0349048. L. X. Zhang’s work was partially supported by grants from the National Science Foundation of China (Project no 10771192).

REFERENCES


W. F. Rosenberger (2002). Randomized urn models and sequential design (with discussion). *Sequential Analysis*, 21, 1–41.


Received 15 November 2006
Accepted 27 July 2008

Feifang HU: fh6e@virginia.edu
Department of Statistics
University of Virginia
Halsey Hall, Charlottesville
Virginia 22904-4135, USA

Li-Xin ZHANG: stazlx@zju.edu.cn
Department of Mathematics
Zhejiang University
Hangzhou 310028, PR China

Siu H. CHEUNG: shcheung@sta.cuhk.edu.hk
Department of Statistics
The Chinese University of Hong Kong
Shatin, Hong Kong, PR China
and
Department of Statistics
National Cheng Kung University
Tainan, Taiwan

Wai S. CHAN: chanws@cuhk.edu.hk
Department of Finance
The Chinese University of Hong Kong
Shatin, Hong Kong, PR China