The Annals of Statistics
2012, Vol. 40, No. 1, 30–44
DOI: 10.1214/11-AOS941

SEQUENTIAL MONITORING WITH CONDITIONAL RANDOMIZATION TESTS

BY VICTORIA PLAMADEALA AND WILLIAM F. ROSENBERGER[1]

Precision Therapeutics and George Mason University

Sequential monitoring in clinical trials is often employed to allow for early stopping and other interim decisions, while maintaining the type I error rate. However, sequential monitoring is typically described only in the context of a population model. We describe a computational method to implement sequential monitoring in a randomization-based context. In particular, we discuss a new technique for the computation of approximate conditional tests following restricted randomization procedures and then apply this technique to approximate the joint distribution of sequentially computed conditional randomization tests. We also describe the computation of a randomization-based analog of the information fraction. We apply these techniques to a restricted randomization procedure, Efron's [*Biometrika* **58** (1971) 403–417] biased coin design. These techniques require derivation of certain conditional probabilities and conditional covariances of the randomization procedure. We employ combinatoric techniques to derive these for the biased coin design.

1. Introduction. Sequential monitoring refers to analyzing data periodically during the course of a clinical trial, with the purpose of detecting early evidence in support of or against a hypothesis. A desirable feature of such a monitoring plan would be flexible inspections of the data that can occur at arbitrary time points. At the same time, sequentially tested hypotheses must maintain the overall probability of type I error at the prespecified level, since repeated testing is known to inflate it. The Lan and DeMets (1983) error spending approach for sequential monitoring allows this. The approach makes use of a type I error spending function, which depends on the amount of "statistical information" available at the time of the interim inspection. In the context of sequential monitoring, the statistical information is a measure of how far a trial has progressed. Under a population model, the amount of interim information—the information fraction—is defined as the proportion of Fisher's information observed thus far in the trial. The type I error spending function rations the amount of type I error that may be spent at each look commensurate to the information fraction. The critical value associated with the allowable probability of type I error at a certain interim look is obtained and

Received July 2011; revised November 2011.

[1]Supported by NSF Grant DMS-09-04253 under the 2009 American Reinvestment and Recovery Act.

MSC2010 subject classifications. Primary 62E15, 62K99; secondary 62L05, 62J10.

Key words and phrases. Biased coin design, conditional reference set, random walk, restricted randomization, sequential analysis.

compared to the observed value of the statistic. The decision whether to continue, or stop, the trial is based on this comparison. Sequential monitoring is typically discussed in the context of a population model. However, it is not uncommon for the FDA to require a "re-analysis" of data using a "re-randomization test," or, as we call it here, a *randomization test*, defined below.

Let $\mathbf{T} = T_1, \ldots, T_n$ be a randomization sequence, where $T_i = 1$ if subject i is randomized to treatment 1; $T_i = 0$ if subject i is randomized to treatment 2, $i = 1, \ldots, n$. Let $N_1(j) = \sum_{i=1}^{j} T_i$ be the number of subjects randomized to treatment 1 after j assignments. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be the responses based on some primary outcome variable, and let $\mathbf{x}$ be the realization. A valid test of the treatment effect can be conducted permuting $\mathbf{T}$ in all possible ways [e.g., Lehmann (1986), Chapter 5]. However, if one wishes to incorporate the randomized design into the analysis, under restricted randomization, such permutations are not equiprobable [e.g., Rosenberger and Lachin (2002), Chapter 7]. The family of *linear rank tests* provides a large class of test statistics with which to conduct randomization tests. The form of the statistic is $V(\mathbf{T}) = \mathbf{a}_n'\mathbf{T}$, for a score vector $\mathbf{a}_n = (a_{1n} - \bar{a}_n, \ldots, a_{nn} - \bar{a}_n)'$, where a_{jn} is some function of the rank of the jth observation and $\bar{a}_n = \sum_{j=1}^{n} a_{jn}/n$. The p-value of the randomization test is computed with respect to a reference set of sequences. The *unconditional reference set* contains all possible allocation sequences, including those that give little or no information about the treatment effect (e.g., $1, 1, \ldots, 1$). Also, the random numbers on each treatment arm, $N_1(n)$ and $N_2(n)$, are ancillary statistics, and therefore the *conditional reference set* is preferred, which finds probabilities conditional on $N_1(n) = n_1$, that is, the observed number on treatment 1 [e.g., Cox (1982), Berger (2000)]. This leads to a *conditional test*.

The literature is largely silent on the subject of sequential monitoring of randomization tests (brief exceptions are found in Rosenberger and Lachin [(2002), Section 7.10] and Zhang and Rosenberger (2008), whose techniques only extend to one interim inspection). The computation of *conditional* randomization tests is also inherently difficult, even without sequential monitoring. We address these issues in this paper by proposing a technique, based on deriving exact conditional distributions of randomization procedures, that leads to a simple computational method for approximating the distribution of sequentially computed randomization tests. We also discuss the appropriate analog for "information fraction" in the context of a randomization model. Our focus will be on one particular restricted randomization procedure, Efron's (1971) biased coin design, which induces a beautiful closed-form combinatoric structure that facilitates such an analysis. However, the technique can be applied to any randomization procedure for which we can determine certain exact conditional distributional results.

Let ϕ_{j+1} be a restricted randomization procedure such that

$$\phi_{j+1} = \Pr\big(T_{j+1} = 1 | N_1(j)\big).$$

Efron's (1971) biased coin design is a restricted randomization procedure for clinical trials that has exceptional properties: it balances treatment assignments throughout the course of the trial with low variability [e.g., Antognini (2008)], and it mitigates selection and accidental biases [Rosenberger and Lachin (2002)]. Then the biased coin design BCD(p) for a parameter $p \in [1/2, 1]$, $q = 1 - p$, is defined as

$$\phi_{j+1} = \begin{cases} 1/2, & \text{when } N_1(j) = j/2, \\ p, & \text{when } N_1(j) < j/2, \\ q, & \text{when } N_1(j) > j/2. \end{cases} \quad j = 0, 1, 2, \ldots, \tag{1.1}$$

Note that $p = 0.5$ results in complete randomization and $p = 1$ results in permuted blocks with block size 2. When $p < 1$, the design is *fully randomized*, in that each subject will be assigned to treatment randomly, which differs markedly from the permuted block design, where some subjects in the tail of each block are assigned with probability 1. Let $D_j = 2N_1(j) - j$ be the difference in numbers assigned to treatments 1 and 2; $\{|D_n|\}_{n=1}^{\infty}$ forms an asymmetric random walk when $p \in (1/2, 1)$. Markaryan and Rosenberger (2010) derive the exact distribution of D_j for the BCD(p), from which the exact distribution of $N_1(n)$ follows immediately:

$$\begin{aligned} &P\big(N_1(n) = n_1\big) \\ &= \begin{cases} p^{n_1} \displaystyle\sum_{l=0}^{n_1-1} \frac{n-2l}{n+2l} \binom{n_1+l}{l} q^l, & n_1 = \dfrac{n}{2}, \\ \dfrac{p^{n_1}}{2} \displaystyle\sum_{l=0}^{n_1} \frac{n-n_1-l}{n-n_1+l} \binom{n-n_1+l}{l} q^{n-2n_1+l-1}, & 0 \le n_1 < \dfrac{n}{2}, \\ \dfrac{p^{n-n_1}}{2} \displaystyle\sum_{l=0}^{n-n_1} \frac{n_1-l}{n_1+l} \binom{n_1+l}{l} q^{2n_1-n+l-1}, & \dfrac{n}{2} < n_1 \le n. \end{cases} \end{aligned} \tag{1.2}$$

Their paper also provides the exact expression for the variance–covariance matrix of the treatment assignments $\mathbf{T}$.

In this paper we provide the exact conditional distribution of $N_1(n)$ given $N_1(j)$, $1 \le j < n$, and the expression for the variance–covariance matrix of $\mathbf{T}$ given $N_1(n)$, $\boldsymbol{\Sigma}_{|n_1}$. While these are heretofore unknown results on theoretical properties of a random walk, our primary interest is that these results give us a computational method to approximate conditional randomization tests following the BCD(p). We then extend these results to the case where sequential analysis is implemented in the course of a clinical trial.

Rosenberger and Lachin (2002) distinguish among three techniques that can be used to compute randomization tests: exact, Monte Carlo and asymptotic. Exact tests are computationally intensive, even with today's computing, and require networking algorithms [Mehta, Patel and Wei (1988)]. Hollander and Peña (1988) developed a clever recursive algorithm to determine the exact distribution of both

conditional and unconditional randomization tests following Efron's biased coin design and applied it to a sample of size of $n = 37$. It can be assumed that such computational techniques would be able to solve much larger problems with today's computing resources. While authors have determined the asymptotic normality of conditional randomization tests under various score functions and randomization procedures [e.g., Smythe (1988)], Efron's biased coin induces a stationary distribution, and hence the test statistic may not be asymptotically normal. This phenomenon was noted in a number of papers, first by counterexample in Smythe and Wei (1983) for the unconditional test, and then by simulation by Hollander and Peña (1988) for the conditional test.

Mehta, Patel and Senchaudhuri (1988) use importance sampling to estimate the conditional randomization test's p-value; their technique employs an elegant, but complex, networking algorithm. The efficiency of the estimator relies on the convergence to normality of the test statistic, which may not hold under the biased coin design. One might be able to modify the network algorithm in Mehta, Patel and Wei (1988) or the recursive algorithm in Hollander and Peña (1988) to compute the exact distribution of sequentially monitored conditional randomization tests, but here we provide a method that is not very computationally intensive and allows us to sample directly from the conditional reference set under a broad class of restricted randomization procedures.

The paper is organized as follows. In Section 2, we present a method for sampling directly from the conditional distribution of $V(\mathbf{T})$, which facilitates the computation of conditional tests. We need to compute certain exact conditional probabilities to apply this method, and we do this for Efron's biased coin design. We extend this application to develop a computational technique for sequential monitoring of conditional randomization tests in Section 3. In Section 4, we describe the analog of "information" in the context of a randomization model. In defining a randomization-based information fraction, we must derive the conditional variance–covariance matrix of $\mathbf{T}$, which we do for Efron's biased coin design. We draw conclusions in Section 5. All the major proofs, some of which require careful combinatorics, are relegated to the online supplement.

2. Computation of conditional randomization tests.

2.1. *Generating sequences from the unconditional reference set.* Suppose a total of n subjects are randomized to two treatments. Let n_1 be the observed number of assignments on treatment 1. A conditional randomization test can be approximated by sampling sufficiently many sequences from the conditional reference set, Ω_c, the collection of sequences that satisfy $N_1(n) = n_1$. This can be achieved by generating sequences from the unconditional reference set, the set of all possible assignments, and retaining those that belong to Ω_c.

Suppose at least N_c number of sequences that satisfy $N_1(n) = n_1$ are sufficient to approximate the conditional randomization distribution of $V(\mathbf{T})$. Let K

sequences be sampled, $\mathbf{T}_1, \ldots, \mathbf{T}_K$, independently and with replacement from the unconditional reference using ϕ_{j+1} as the sampling mechanism. This number of Monte Carlo sequences must be large enough such that at least N_c sequences satisfy the condition $N_1(n) = n_1$. The requisite number of sequences, K, follows a negative binomial random variable with parameters $\pi = P(N_1(n) = n_1)$ and $r = N_c$ [Zhang and Rosenberger (2011)]. Let N denote a value in the range of K, $N = N_c, N_c + 1, \ldots$. For $k = 1, \ldots, N$, a sequence $\mathbf{T}_k = \mathbf{t}$ is sampled from the unconditional reference set with probability

$$f(\mathbf{t}) = (1/2) \prod_{j=1}^{n-1} (\phi_{j+1})^{t_{j+1}} (1 - \phi_{j+1})^{1-t_{j+1}}, \tag{2.1}$$

where t_{j+1} are the observed values of T_{j+1}. The jth sampled sequence induces two Bernoulli random variables

$$Y_j = \begin{cases} 1, & \text{if } N_1(n) = n_1, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$X_j = \begin{cases} 1, & \text{if } N_1(n) = n_1 \text{ and } V(\mathbf{T}_j) \geq v^*, \\ 0, & \text{otherwise,} \end{cases}$$

where v^* is the observed value of the statistic. A strongly consistent estimator for the p-value of the upper-tailed conditional test can be computed as

$$\hat{p}_c = \frac{\sum_{j=1}^{N} X_j}{\sum_{j=1}^{N} Y_j}. \tag{2.2}$$

Table 1 reports the 95th percentile of K when sampling from the unconditional reference set and $N_c = 2500$ for Efron's biased coin design. These sample sizes are reasonable when there is perfect balance in the assignments, but increase considerably in the presence of imbalance. This technique cannot be used in the presence of any imbalance.

TABLE 1
Approximate 95th percentile of K for various n, n_1, $N_c = 2500$

n	$n_1 = 0.45n$	$n_1 = 0.48n$	$n_1 = 0.50n$
	BCD($p = 2/3$)		
100	3,531,344	55,060	5117
200	3,611,280,266	881,557	5117
500	$3{,}877{,}310 \times 10^{12}$	3,611,026,232	5117
	BCD($p = 3/4$)		
100	114,384,212	156,865	3822
200	$6{,}754{,}269 \times 10^{6}$	12,709,307	3822
500	$1{,}390{,}644 \times 10^{21}$	$6{,}754{,}269 \times 10^{6}$	3822

2.2. *Our method*: *Generating sequences from the conditional reference set*. Rather than sampling too many sequences and discarding those that do not satisfy the condition $N_1(n) = n_1$, it is more efficient to sample directly from Ω_c—the collection of all randomization sequences that satisfy the condition $N_1(n) = n_1$. The set Ω_c will be called the conditional reference set. Let N_c randomization sequences, $\mathbf{T}_1, \ldots, \mathbf{T}_{N_c}$, be sampled independently and with replacement strictly from Ω_c, each with respective probabilities $h(\mathbf{t}_1), \ldots, h(\mathbf{t}_{N_c})$. For an upper-tailed test, the kth sampled sequence induces a Bernoulli random variable

$$V_k = \begin{cases} 1, & \text{if } V(\mathbf{T}_k) \geq v^*, \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

The Monte Carlo estimator of the upper-tailed test's p-value is the strongly consistent and unbiased estimator $\bar{V} = \sum_{k=1}^{N_c} V_k / N_c$. (It may be possible to find an estimator with a smaller variance, but we do not address the issue of estimation of p-values in this paper.)

To guarantee a sequence from Ω_c, T_{j+1} in ϕ_{j+1} must be conditioned on both $N_1(j)$ and $N_1(n)$. Consequently, for $0 \leq m_j \leq j$, the procedure

$$p_{j+1} = \begin{cases} P\big(T_{j+1} = 1 | N_1(j) = m_j, N_1(n) = n_1\big), & 1 \leq j \leq n-1, \\ P\big(T_{j+1} = 1 | N_1(n) = n_1\big), & j = 0, \end{cases} \tag{2.4}$$

must be applied to generate a random sequence strictly from Ω_c. We now provide a general formula relating the conditional and the unconditional reference sets, which facilitates the generation of sequences from the conditional reference set for any restricted randomization procedure of the form $\phi_{j+1} = \Pr(T_{j+1} = 1 | N_1(j))$.

THEOREM 2.1. *For* $n = 1, 2, 3, \ldots, 0 \leq n_1 \leq n$, $0 \leq j < n$, $0 \leq m_j \leq j$ *and* $\phi_{j+1}(m_j) = P(T_{j+1} = 1 | N_1(j) = m_j)$, *the rule*

$$p_{j+1} = \begin{cases} \phi_{j+1}(m_j) \dfrac{P(N_1(n) = n_1 | N_1(j+1) = m_j + 1)}{P(N_1(n) = n_1 | N_1(j) = m_j)}, & \\ & 1 \leq j \leq n-1, \\ \phi_{j+1}(m_j) \dfrac{P(N_1(n) = n_1 | T_{j+1} = 1)}{P(N_1(n) = n_1)}, & j = 0, \end{cases} \tag{2.5}$$

can be used to sample a sequence that satisfies $N_1(n) = n_1$.

PROOF. The result follows from an application of Bayes theorem to (2.4) and the Markovian property of N_1. □

Furthermore, for $k = 1, \ldots, N_c$, a sequence $\mathbf{T}_k = \mathbf{t}$ is sampled from Ω_c with probability

$$h(\mathbf{t}) = \prod_{j=0}^{n-1} (p_{j+1})^{t_{j+1}} (1 - p_{j+1})^{1 - t_{j+1}}. \tag{2.6}$$

In the simplest case, complete randomization, $p_{j+1} = (n_1 - m_j)/(n-j), 0 \le j \le n-1$, and this is the random allocation rule [see Rosenberger and Lachin (2002)], which is sometimes used to fill permuted blocks.

The following theorem gives these probabilities for Efron's biased coin design. The distribution of $N_1(n)$ given $N_1(j) = m_j$, $0 \le m_j \le j$, has three cases depending on the value of m_j with respect to j, $1 \le j < n$. Within each case, $P(N_1(n) = n_1 | N_1(j) = m_j)$ depends the value of n_1 with respect to n, j and m_j.

THEOREM 2.2. *Let $n = 2, 3, 4, \ldots, 1 \le j < n$, $0 \le m_j \le j$ and $m_j \le n_1 \le n - j + m_j$. Denote*

$$C(x,l) := \frac{x-l}{x+l}\binom{x+l}{l} \quad \text{and} \quad D := \binom{n-j}{n_1 - m_j} - \binom{n-j}{n_1 - j + m_j}.$$

For the BCD(p):

(1) *When $0 \le m_j < j/2$, $P(N_1(n) = n_1 | N_1(j) = m_j)$ is*

$$\begin{cases} \binom{n-j}{n_1 - m_j} p^{n_1 - m_j} q^{n-j-n_1+m_j} & \text{if } m_j \le n_1 < j - m_j, \\ 0.5 p^{n_1 - m_j} \displaystyle\sum_{l=0}^{n_1+m_j-j} C(n - n_1 - m_j, l) q^{n-2n_1-1+l} & \\ \quad + D p^{n_1-m_j} q^{n-j-n_1+m_j} & \text{if } j - m_j \le n_1 < n/2, \\ p^{n_1-m_j} \displaystyle\sum_{l=0}^{n-j-n_1+m_j} C(n_1 - m_j, l) q^l & \text{if } n_1 = n/2, \\ 0.5 p^{n-n_1-m_j} \displaystyle\sum_{l=0}^{n-j-n_1+m_j} C(n_1 - m_j, l) q^{2n_1-n-1+l} & \\ & \text{if } n/2 < n_1 \le n - j + m_j. \end{cases}$$

(2) *When $m_j = j/2$,*

$$\begin{aligned} &P\big(N_1(n) = n_1 | N_1(j) = m_j\big) \\ &\quad = P\big(N_1(n-j) = n_1 - m_j\big), \qquad m_j \le n_1 \le n - j + m_j, \end{aligned}$$

where the unconditional distribution is derived in Markaryan and Rosenberger (2010).

(3) *When $j/2 < m_j \le j$, $P(N_1(n) = n_1 | N_1(j) = m_j)$ is*

$$\begin{cases} 0.5 p^{n_1+m_j-j} \displaystyle\sum_{l=0}^{n_1-m_j} C(n - j - n_1 + m_j, l) q^{n-2n_1-1+l} & \text{if } m_j \le n_1 < n/2, \end{cases}$$

$$p^{n-j-n_1+m_j} \sum_{l=0}^{n_1-m_j} C(n-j-n_1+m_j, l) q^l \qquad \text{if } n_1 = n/2,$$

$$0.5 p^{n-j-n_1+m_j} \sum_{l=0}^{n-n_1-m_j} C(n_1+m_j-j, l) q^{2n_1-n-1+l}$$

$$+ D p^{n-j-n_1+m_j} q^{n_1-m_j} \qquad \text{if } n/2 < n_1 \le n-m_j,$$

$$\binom{n-j}{n_1-m_j} p^{n-j-n_1+m_j} q^{n_1-m_j} \qquad \text{if } n-m_j < n_1 \le n-j+m_j.$$

PROOF. See Appendix A in the supplementary material [Plamadeala and Rosenberger (2011)]. □

Note that if $n = j$ and $n_1 = m_j$, $P(N_1(n) = n_1 | N_1(j) = m_j) = 1$, and if $m_j > n_1$ or $n - j < n_1 - m_j$, $P(N_1(n) = n_1 | N_1(j) = m_j) = 0$. Also, $P(N_1(n) = n_1 | N_1(0) = 0) = P(N_1(n) = n_1)$.

The procedure then follows by simply generating N_c sequences using (2.5). This allows us to reduce the magnitude of the problem from the astronomical numbers in Table 1 to just N_c. A satisfactory value for N_c can be obtained from the constraint $\text{MSE}(\bar{V}) = p_c(1-p_c)/N_c \le 1/4N_c \le \varepsilon$. For $\varepsilon = 0.0001$, $N_c \ge 2500$. Higher precision in estimation is possible by finding N_c that ensures $P(|\bar{V} - p_c| \le 0.1 p_c) = 0.99$, for instance. It follows that $N_c \approx (2.576/0.1)^2(1-p_c)/p_c$. Thus, to estimate a p-value as large as 0.04 with an error of 10% of 0.04 with 0.99 probability, the Monte Carlo sample size must be $N_c = 15{,}924$. If a smaller p-value is expected, N_c will be larger.

Table 2 provides approximations for the upper 0.1 tail of the linear rank statistic with simple rank scores under the BCD(0.6) randomization. For small samples sizes, we also provide the exact p-value for comparison purposes; Monte Carlo

TABLE 2
Approximations for the upper 0.1 *tail of the randomization distribution of the linear rank statistic by sampling from* Ω_c; $N_c = 2500$, BCD(0.6)

	n	n_1	Exact	1000 Monte Carlo runs; mean (SD)
$P(V(\mathbf{T}) \ge 21.5)$	30	15	0.1057	0.1053 (0.0061)
$P(V(\mathbf{T}) \ge 23)$	30	12	0.1009	0.1008 (0.0059)
$P(V(\mathbf{T}) \ge 31)$	40	20	0.1011	0.1009 (0.0061)
$P(V(\mathbf{T}) \ge 34)$	40	16	0.1000	0.0997 (0.0060)
$P(V(\mathbf{T}) \ge 82)$	100	50		0.1055 (0.0060)
$P(V(\mathbf{T}) \ge 113)$	100	40		0.1043 (0.0062)
$P(V(\mathbf{T}) \ge 299)$	500	250		0.1104 (0.0063)
$P(V(\mathbf{T}) \ge 1000)$	500	200		0.1030 (0.0058)

estimates are very close with small variability. As expected, the variability of the estimates does not change across different sample sizes n. The computational complexity of the sampling scheme for the BCD is invariant to the value of p. Comparing Tables 1 and 2, the conditional distribution method reduces the Monte Carlo sample size to a few thousand.

Following stratification on known covariates, the computation of a stratified linear rank test based on the conditional randomization distribution is straightforward by summing the stratum-specific linear-rank test statistics over I independent strata. Using the methodology described in this section, a sequence is sampled independently from the conditional reference of each stratum; the linear-rank statistic is evaluated in each stratum and the stratum-specific test statistics are summed. The process is repeated N_c times, and the stratified test's p-value is estimated by the proportion of summed statistics as or more extreme than the one observed.

3. Extension to sequential monitoring. Suppose there are $L-1$ interim inspections of the data after $1 \le r_1 < r_2 < \cdots < r_{L-1} < r_L = n$ patients responded. Let $0 < t_1 < t_2 < \cdots < t_{L-1} < t_L = 1$ be the corresponding information fraction at those inspections. For conditional tests, let $N_1(r_1), N_1(r_2), \ldots, N_1(r_{L-1})$, and $N_1(r_L) = N_1(n)$ be the sample sizes randomized to treatment 1 after inspections $1, \ldots, L$ and let $n_{11}, \ldots, n_{1(L-1)}$, and $n_{1L} = n_1$ be realizations of these sample sizes. Let the linear-rank randomization test statistic computed at each of the inspections be given by $V_{r_l} = \sum_{j=1}^{r_l} (a_{jr_l} - \bar{a}_{r_l}) T_j = \mathbf{a}'_{r_l} \mathbf{T}^{(r_l)}$, $l = 1, \ldots, L$. Using the alpha-spending function approach [Lan and DeMets (1983)], let $\alpha^*(t), t \in [0,1]$, be a nondecreasing function such that $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$, the significance level of the one-sided test. One such function is $\alpha^*(t) = 2 - 2\Phi(z_{\alpha/2}/\sqrt{t})$, $0 < t \le 1$; $\alpha^*(0) = 0$, where Φ is the standard normal distribution function and $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ [Lan and DeMets (1983), O'Brien and Fleming (1979)]. Following Zhang and Rosenberger (2008), the upper-tailed, conditional randomization test with L interim looks involves finding $d_1, \ldots, d_L$ such that

$$
\begin{cases}
P\big(V_{r_1} > d_1 | N_1(r_1) = n_{11}\big) = \alpha^*(t_1), \\
P\big(V_{r_1} \le d_1, V_{r_2} > d_2 | N_1(r_1) = n_{11}, N_1(r_2) = n_{12}\big) = \alpha^*(t_2) - \alpha^*(t_1), \\
P\left(V_{r_1} \le d_1, V_{r_2} \le d_2, V_{r_3} > d_3 \Big| \bigcap_{j=1}^{3} N_1(r_j) = n_{1j}\right) = \alpha^*(t_3) - \alpha^*(t_2), \\
\vdots \\
P\left(V_{r_1} \le d_1, \ldots, V_L > d_L \Big| \bigcap_{j=1}^{L} N_1(r_j) = n_{1j}\right) = \alpha - \alpha^*(t_{L-1}).
\end{cases} \tag{3.1}
$$

The asymptotic joint normality of these conditional distributions has not been shown, except in the case of $L = 2$ under the generalized biased coin design [Zhang and Rosenberger (2008)].

We express (3.1) in terms of univariate conditional distributions, which are much easier to sample from than the joint distributions in (3.1).

LEMMA 3.1. *The set of conditions* (3.1) *is equivalent to*

$$
\begin{cases}
P\big(V_{r_1} > d_1 | N_1(r_1) = n_{11}\big) = \alpha^*(t_1), \\
P\left(V_{r_2} > d_2 | V_{r_1} \le d_1, \bigcap_{j=1}^{2} \{N_1(r_j) = n_{1j}\}\right) = \dfrac{\alpha^*(t_2) - \alpha^*(t_1)}{1 - \alpha^*(t_1)}, \\
P\left(V_{r_3} > d_3 \Big| \bigcap_{j=1}^{2} \{V_{r_j} \le d_j\}, \bigcap_{j=1}^{3} \{N_1(r_j) = n_{1j}\}\right) = \dfrac{\alpha^*(t_3) - \alpha^*(t_2)}{1 - \alpha^*(t_2)}, \\
\vdots \\
P\left(V_n > d_L \Big| \bigcap_{j=1}^{L-1} \{V_{r_j} \le d_j\}, \bigcap_{j=1}^{L} \{N_1(r_j) = n_{1j}\}\right) = \dfrac{\alpha - \alpha^*(t_{L-1})}{1 - \alpha^*(t_{L-1})}.
\end{cases}
\tag{3.2}
$$

PROOF. See Appendix B in the supplementary material [Plamadeala and Rosenberger (2011)]. □

At each inspection l in (3.2), the conditional reference set is the collection of all sequences satisfying $\bigcap_{i=1}^{l} \{N_1(r_i) = n_{1i}\}$. The following theorem can be used to sample sequences from such sets.

THEOREM 3.1. *Let* $1 \le l \le L$, $r_0, r_1, r_2, \ldots, r_l$ *and* $n_{10}, n_{11}, \ldots, n_{1l}$ *be defined as before, with* $r_0 = 0$ *and* $n_{10} = 0$. *Let* $k = 1, \ldots, l$. *For* $r_{k-1} \le j < r_k$, $n_{1(k-1)} \le m_j \le j$ *and* $\phi_{j+1}(m_j) = P(T_{j+1} = 1 | N_1(j) = m_j)$, *the rule*

$$
\psi_{j+1} = \phi_{j+1}(m_j) \frac{P(N_1(r_k) = n_{1k} | N_1(j+1) = m_j + 1)}{P(N_1(r_k) = n_{1k} | N_1(j) = m_j)} \tag{3.3}
$$

can be used to sample a sequence that satisfies $\bigcap_{i=1}^{l} \{N_1(r_i) = n_{1i}\}$.

PROOF. See Appendix C in the supplementary material [Plamadeala and Rosenberger (2011)]. □

Note that equation (3.3) reduces succinctly to the expected $\psi_{j+1} = (n_{1k} - m_j)/(r_k - j)$ for complete randomization, $l = 1, \ldots, L$, $k = 1, \ldots, l$, $r_{k-1} \le j < r_k$ and $n_{1(k-1)} \le m_j \le j$. For the BCD(p) the numerator and the denominator of ψ_{j+1} must be evaluated according to Theorem 2.2. To obtain a sequence from the reference set satisfying $\bigcap_{i=1}^{l} \{N_1(r_i) = n_{1i}\}$, the sampling must be done in $k = 1, \ldots, l$ steps as follows:

(1) At stage $k = 1$, apply ψ_{j+1} with $r_0 \le j < r_1$ to sample the first r_1 assignments.

(2) At stage $k = 2$, apply ψ_{j+1} with $r_1 \le j < r_2$ to sample the next $r_2 - r_1$ assignments.

(3) At stage $3 \le k \le l$, apply ψ_{j+1} with $r_{k-1} \le j < r_k$ to sample the next $r_k - r_{k-1}$ assignments.

Suppose a sample of size N_c (sequences) is sufficient to estimate a distribution quantile using some quantile estimator. The Monte Carlo algorithm that estimates the boundary $d_1, \ldots, d_L$ for an α-level, upper-tailed, conditional randomization test with $L - 1$ interim inspections is as follows:

(1) At stage 1, generate N_c randomization sequences of r_1 assignments from the reference set satisfying $N_1(r_1) = n_{11}$. Evaluate V_{r_1} for each sequence; estimate d_1 using the nonparametric quantile estimator of Chen and Lazar (2010) based on the values of V_{r_1}.

(2) At stage 2, generate $N_c/(1-\alpha^*(t_1))$ randomization sequences of r_2 assignments from the reference set satisfying $\bigcap_{i=1}^{2}\{N_1(r_i) = n_{1i}\}$. For each sequence, evaluate V_{r_1} using the first r_1 of r_2 assignments only. Retain those sequences that satisfy $\{V_{r_1} \le d_1\}$. Evaluate V_{r_2} for each retained sequence. Estimate d_2 using the quantile estimator of Chen and Lazar (2010) based on the values of V_{r_2}.

(3) At stage $3 \le l \le L$, generate $N_c/\prod_{i=1}^{l-1}(1 - [\alpha^*(t_i) - \alpha^*(t_{i-1})]/[1 - \alpha^*(t_{i-1})])$ randomization sequences of r_l assignments from the reference set satisfying $\bigcap_{i=1}^{l}\{N_1(r_i) = n_{1i}\}$. Note that $\alpha^*(t_0) = 0$ and $\alpha^*(t_L) = \alpha$. For each sequence, evaluate $V_{r_1}, V_{r_2}, \ldots, V_{r_{l-1}}$ using the first $r_1, r_2, \ldots, r_{l-1}$ assignments, respectively. Retain those sequences that satisfy $\bigcap_{i=1}^{l-1}\{V_{r_i} \le d_i\}$. Evaluate V_{r_l} for each retained sequence. Estimate d_l using the quantile estimator of Chen and Lazar (2010) based on the values of V_{r_l}.

Requiring that $N_c/\prod_{i=1}^{l-1}(1 - [\alpha^*(t_i) - \alpha^*(t_{i-1})]/[1 - \alpha^*(t_{i-1})])$ randomization sequences be sampled at stage l simply ensures that at least N_c sequences are used for the estimation of d_l at each stage l.

4. Randomization-based information. Fisher's information is defined under a population model, and hence it is not defined in the context of randomization-based inference. However, since the Fisher's information approximates the inverse of the asymptotic variance of the test, it seems reasonable to define the randomization-based analog of information as the ratio of the variances [Rosenberger and Lachin (2002)].

$$t_l = \frac{\mathbf{a}'_{r_l}\boldsymbol{\Sigma}_{|r_l}\mathbf{a}_{r_l}}{\mathbf{a}'_n\boldsymbol{\Sigma}_{|n}\mathbf{a}_n}, \tag{4.1}$$

where $\boldsymbol{\Sigma}_{|r_l} = \mathrm{Var}(\mathbf{T}^{(r_l)}|N_1(r_1) = n_{11}, \ldots, N_1(r_l) = n_{1l})$. This requires specification of $\boldsymbol{\Sigma}_{|r_l}$ and $\boldsymbol{\Sigma}_{|n}$. We now derive these for Efron's biased coin design. We begin with three lemmas:

LEMMA 4.1. *Let $n = 2, 3, \ldots$ and $0 \le n_1 \le n$. Let $\phi_i(a) = P(T_i = 1|N_1(i-1) = a)$ and $f_{j-1,b}^{(i,a+1)} = P(N_1(j-1) = b|N_1(i) = a+1)$. For $1 \le i < j \le n$,*

$$E\big(T_i T_j | N_1(n) = n_1\big) = \frac{\sum_{a=0}^{i-1} \phi_i(a) P(N_1(i-1) = a) \sum_{b=a+1}^{j-1} \phi_j(b) f_{j-1,b}^{(i,a+1)} f_{n,n_1}^{(j,b+1)}}{P(N_1(n) = n_1)}.$$

The conditional probabilities $f_{j-1,b}^{(i,a+1)}$ and $f_{n,n_1}^{(j,b+1)}$ are given by Theorem 2.2.

PROOF. The result follows from an application of Bayes theorem to $P(T_i = 1, T_j = 1|N_1(n) = n_1)$ and the Markovian property of N_1. □

Given that we observe $N_1(n) = n_1$, we now derive the variance–covariance matrix of $\mathbf{T}$, denoted by $\mathbf{\Sigma}_{|n_1}$.

LEMMA 4.2. *Let $n = 1, 2, \ldots, 0 \le n_1 \le n$, $\vartheta_{i|n_1} = E(T_i|N_1(n) = n_1)$ and $\phi_i(a) = P(T_i = 1|N_1(i-1) = a)$. For the* BCD($p$)

$$\vartheta_{i|n_1} = \frac{\sum_{a=0}^{i-1} P(N_1(i-1) = a)\phi_i(a) P(N_1(n) = n_1|N_1(i) = a+1)}{P(N_1(n) = n_1)},$$

where

$$\vartheta_{1|n_1} = 1/2 P\big(N_1(n) = n_1|N_1(1) = 1\big)/P\big(N_1(n) = n_1\big).$$

If $i < j$, the (i, j)th entry of $\mathbf{\Sigma}_{|n_1}$ is

$$\sigma_{ij} = \frac{\sum_{a=0}^{i-1} \phi_i(a) P(N_1(i-1) = a) \sum_{b=a+1}^{j-1} \phi_j(b) f_{j-1,b}^{(i,a+1)} f_{n,n_1}^{(j,b+1)}}{P(N_1(n) = n_1)} - \vartheta_{i|n_1}\vartheta_{j|n_1}.$$

If $i = j$, the (i, j)th entry of $\mathbf{\Sigma}_{|n_1}$ is

$$\sigma_{ij} = \vartheta_{i|n_1}(1 - \vartheta_{i|n_1}).$$

PROOF. The result follows from an application of Bayes theorem to $P(T_i = 1|N_1(n) = n_1)$, the Markovian property of N_1 and Lemma 4.1. □

LEMMA 4.3. *Let $1 \le l \le L$, $r_0, r_1, r_2, \ldots, r_l$ and $n_{10}, n_{11}, \ldots, n_{1l}$ be defined as before, with $r_0 = n_{10} = 0$. Let $\phi_i(a) = P(T_i = 1|N_1(i-1) = a)$, $k = 1, \ldots, l$, and $f_{i-1,a}^{(r_{k-1}, n_{1(k-1)})} = P(N_1(i-1) = a|N_1(r_{k-1}) = n_{1(k-1)})$. Denote $\vartheta_{i|r_l} = E(T_i|\bigcap_{q=1}^{l}\{N_1(r_q) = n_{1q}\})$ and $\lambda_{ij|r_l} = E(T_i T_j|\bigcap_{q=1}^{l}\{N_1(r_q) = n_{1q}\})$.*

For $1 \le k \le l$, $r_{k-1} < i \le r_k$,

$$\vartheta_{i|r_l} = \frac{\sum_{a=n_{1(k-1)}}^{i-1} \phi_i(a) f_{i-1,a}^{(r_{k-1}, n_{1(k-1)})} f_{r_k,n_{1k}}^{(i,a+1)}}{P(N_1(r_k) = n_{1k}|N_1(r_{k-1}) = n_{1(k-1)})}.$$

For $1 \le k \le l$ *and* $r_{k-1} < i < j \le r_k$,

$$\lambda_{ij|r_l} = \frac{\sum_{a=n_{1(k-1)}}^{i-1} \phi_i(a) f_{i-1,a}^{(r_{k-1},n_{1(k-1)})} \sum_{b=n_{1(k-1)+1}}^{j-1} \phi_j(b) f_{j-1,b}^{(i,a+1)} f_{r_k,n_{1k}}^{(j,b+1)}}{f_{r_k,n_{1k}}^{(r_{k-1},n_{1(k-1)})}}.$$

For all other i, j,

$$\lambda_{ij|r_l} = E\left(T_i \Big| \bigcap_{q=1}^{l} \{N_1(r_q) = n_{1q}\}\right) E\left(T_j \Big| \bigcap_{q=1}^{l} \{N_1(r_q) = n_{1q}\}\right).$$

The probabilities $f_{i-1,a}^{(r_{k-1},n_{1(k-1)})}$, $f_{r_k,n_{1k}}^{(i,a+1)}$, $f_{j-1,b}^{(i,a+1)}$, $f_{r_k,n_{1k}}^{(j,b+1)}$ *and* $f_{r_k,n_{1k}}^{(r_{k-1},n_{1(k-1)})}$ *are given by Theorem* 2.2.

PROOF. See Appendix D in the supplementary material [Plamadeala and Rosenberger (2011)]. □

Finally, the closed form of $\mathbf{\Sigma}_{|r_l}$ is given in the following theorem, which follows immediately from Lemma 4.3:

THEOREM 4.1. *Let* $1 \le l \le L, k = 1, \ldots, l, r_0, r_1, r_2, \ldots, r_l$ *and* $n_{10}, n_{11}, \ldots, n_{1l}$ *be defined as before, with* $r_0 = n_{10} = 0$.

The (i, j)*th entry of* $\mathbf{\Sigma}_{|r_l}$ *under the* BCD(p) *is*

$$\sigma_{ij} = \begin{cases} \lambda_{ij|r_l} - \vartheta_{i|r_l}\vartheta_{j|r_l}, & \text{if } i < j \text{ and } r_{k-1} < i < j \le r_k, \\ \vartheta_{i|r_l}(1 - \vartheta_{i|r_l}), & \text{if } i = j, \\ 0, & \text{otherwise}, \end{cases}$$

where $\vartheta_{i|r_l}$ *and* $\lambda_{ij|r_l}$ *are given by Lemma* 4.3.

Although one can compute $\mathbf{\Sigma}_{|n}$ and $\mathbf{\Sigma}_{|r_l}$ exactly using Theorem 4.1, $\mathbf{a}'_n$ in (4.1) remains unknown at each interim inspection, since a portion of the data is unobserved. One would have to interpolate sequentially the remaining unknown data points in order to have a value for $\mathbf{a}'_n$ and an approximation for (4.1). Interpolating the unknown observations by sampling with replacement the known observations is one way to obtain a value for $\mathbf{a}'_n$. In our simulations with data generated from two normal distributions, $L = 3$, $n = 350$, $n_1 = 174$ and assignments following the BCD(3/4), the approximate information fraction at the first interim look with $r_1 = 250$ and $n_{11} = 126$ was 0.3791, compared to the true information of 0.3759. At the second interim look with $r_2 = 300$ and $n_{12} = 148$, the approximate information fraction was 0.6380, compared to the true information of 0.6382.

We also simulate the probability of type I error in an example. For this purpose, we generate a sample of $n = 350$ observations from $N(1, 0.9)$ and simulate treatment assignments from BCD($p = 3/4$). We plan $L = 3$ interim looks: at $r_1 = 250$, $r_2 = 300$ and $r_3 = 350$. The observed number assigned to treatment 1 at each look

TABLE 3

Mean (SD) of simulated α for an α = 0.05 upper tail sequential test over a Monte Carlo sample size of 1000, $N_c = 2500$, interpolating the unknown observations by sampling with replacement

Look l	r_l	n_{1l}	t_l	$\alpha_l{}^*$	$\hat{d}_l$	$\hat{\alpha}$
Look 1	250	126	0.3617	0.0011	1709	
Look 2	300	148	0.6248	0.0121	1688	
Look 3	350	174	1	0.0373	1501	0.0495 (0.0043)

$^*\alpha_l = \frac{\alpha^*(t_l) - \alpha^*(t_{l-1})}{1 - \alpha^*(t_{l-1})}$.

was $n_{11} = 126$, $n_{12} = 128$ and $n_{13} = 174$. We compute the boundary values using the algorithm in Section 3. Table 3 gives the estimated type I error rate ($\hat{\alpha}$) and standard deviation over 1000 replications for this sequential conditional test. The probability of type I error is preserved with low variability.

5. Conclusions. We have provided a computational method to approximate conditional randomization tests, which can be extended to clinical trials that incorporate sequential monitoring. The key is to determine certain conditional probabilities from the particular randomization procedure. These techniques apply to any restricted randomization procedure of the form $\phi_{j+1} = \Pr(T_{j+1} = 1 | N_1(j))$ and for which closed form conditional probabilities can be obtained. We have derived the exact conditional distribution of $N_1(n)$, given $N_1(j)$, for Efron's BCD(p) using combinatoric arguments, also the conditional variance–covariance matrix of $\mathbf{T}$, which allows computation of the information fraction.

The class of generalized biased coin designs (GBCD) [Wei (1978)] does not have a known form for the exact conditional distribution, and this remains an open problem. For the sequential monitoring of conditional tests using the GBCD with one interim look, Zhang and Rosenberger (2008) derived the joint asymptotic distribution of the interim and the final test statistics, which allows for an asymptotic test.

Acknowledgments. The authors thank Tigran Markaryan, Anindya Roy and the referees for helpful comments.

SUPPLEMENTARY MATERIAL

Supplement to "Sequential monitoring with conditional randomization tests" (DOI: 10.1214/11-AOS941SUPP; .pdf). The supplement contains Appendix A (proof of Theorem 2.2), Appendix B (proof of Lemma 3.1), Appendix C (proof of Theorem 3.1), and Appendix D (proof of Lemma 4.3).

REFERENCES

ANTOGNINI, A. B. (2008). A theoretical analysis of the power of biased coin designs. *J. Statist. Plann. Inference* **138** 1792–1798. MR2400479

BERGER, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Stat. Med.* **19** 1319–1328.

CHEN, J. and LAZAR, N. A. (2010). Quantile estimation for discrete data via empirical likelihood. *J. Nonparametr. Stat.* **22** 237–255. MR2599016

COX, D. R. (1982). A remark on randomization in clinical trials. *Util. Math.* **21A** 242–252.

EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417. MR0312660

HOLLANDER, M. and PEÑA, E. (1988). Nonparametric tests under restricted treatment-assignment rules. *J. Amer. Statist. Assoc.* **83** 1144–1151. MR0997593

LAN, K. K. G. and DEMETS, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70** 659–663. MR0725380

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York. MR0852406

MARKARYAN, T. and ROSENBERGER, W. F. (2010). Exact properties of Efron's biased coin randomization procedure. *Ann. Statist.* **38** 1546–1567. MR2662351

MEHTA, C. R., PATEL, N. R. and SENCHAUDHURI, P. (1988). Importance sampling for estimating exact probabilities in permutational inference. *J. Amer. Statist. Assoc.* **83** 999–1005. MR0997575

MEHTA, C. R., PATEL, N. R. and WEI, L. J. (1988). Constructing exact significance tests with restricted randomization rules. *Biometrika* **75** 295–302.

O'BRIEN, P. C. and FLEMING, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35** 549–556.

PLAMADEALA, V. and ROSENBERGER, W. F. (2011). Supplement to "Sequential monitoring with conditional randomization tests." DOI:10.1214/11-AOS941SUPP.

ROSENBERGER, W. F. and LACHIN, J. M. (2002). *Randomization in Clinical Trials*: *Theory and Practice*. Wiley, New York. MR1914364

SMYTHE, R. T. (1988). Conditional inference for restricted randomization designs. *Ann. Statist.* **16** 1155–1161. MR0959193

SMYTHE, R. T. and WEI, L. J. (1983). Significance tests with restricted randomization design. *Biometrika* **70** 496–500. MR0712039

WEI, L. J. (1978). The adaptive biased coin design for sequential experiments. *Ann. Statist.* **6** 92–100. MR0471205

ZHANG, Y. and ROSENBERGER, W. F. (2008). Sequential monitoring of conditional randomization tests: Generalized biased coin designs. *Sequential Anal.* **27** 234–253. MR2446901

ZHANG, L. and ROSENBERGER, W. F. (2011). Adaptive randomization in clinical trials. In *Design and Analysis of Experiments*, *Vol.* III. (K. Hinkelmann, ed.). Wiley, Hoboken.

PRECISION THERAPEUTICS
2516 JANE STREET
PITTSBURGH, PENNSYLVANIA 15203
USA
E-MAIL: vplamadeala@ptilabs.com

DEPARTMENT OF STATISTICS
GEORGE MASON UNIVERSITY
4400 UNIVERSITY DRIVE, MS 4A7
FAIRFAX, VIRGINIA 22030
USA
E-MAIL: wrosenbe@gmu.edu