



Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization

Richard Simon ^{a,*}, Noah Robin Simon ^b^a Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, United States^b Department of Statistics, Stanford University, Stanford, CA 94305, United States

ARTICLE INFO

Article history:

Available online xxxx

Keywords:

Response adaptive randomization
Adaptive stratification
Clinical trials

ABSTRACT

We demonstrate that clinical trials using response adaptive randomized treatment assignment rules are subject to substantial bias if there are time trends in unknown prognostic factors and standard methods of analysis are used. We develop a general class of randomization tests based on generating the null distribution of a general test statistic by repeating the adaptive randomized treatment assignment rule holding fixed the sequence of outcome values and covariate vectors actually observed in the trial. We develop broad conditions on the adaptive randomization method and the stochastic mechanism by which outcomes and covariate vectors are sampled that ensure that the type I error is controlled at the level of the randomization test. These conditions ensure that the use of the randomization test protects the type I error against time trends that are independent of the treatment assignments. Under some conditions in which the prognosis of future patients is determined by knowledge of the current randomization weights, the type I error is not strictly protected. We show that response adaptive randomization can result in substantial reduction in statistical power when the type I error is preserved. Our results also ensure that type I error is controlled at the level of the randomization test for adaptive stratification designs used for balancing covariates.

Published by Elsevier B.V.

1. Introduction

There is considerable interest in medicine today in the use of adaptive methods in clinical trials. Adaptive methods can be used in several ways, including adapting randomization weights based on response data. Traditionally a randomized clinical trial assigns the treatments being studied to patients using a fixed weighting scheme. Most commonly with two treatments, the treatments are assigned with equal probability throughout the course of the study. There has long been interest in adjusting the weights as information accumulates about the relative effectiveness of the treatments (Armitage, 1985; Hoel et al., 1975; Rosenberger and Lachin, 1993; Simon, 1977; Zelen, 1969). Such methods have been used infrequently in phase 3 clinical trials, however, because of concern that changing the randomization weights in a data determined manner would impair the validity of the analysis of the results. Of particular concern is the possibility that the false positive error rate might exceed the apparent significance level of the statistical test. Today, there is such interest in the general use of adaptive methods in clinical trials that the US Food and Drug Administration has issued a draft guidance on adaptive methods (FDA, 2010). Our purpose here is to introduce a statistical significance test that preserves the claimed type I error for a very general class of adaptive randomization methods.

* Corresponding author. Tel.: +1 301 496 0975; fax: +1 301 402 0560.

E-mail addresses: rsimon@nih.gov (R. Simon), nsimon@stanford.edu (N.R. Simon).

Table 1
Type I error for Mann–Whitney test.

	No time trend	Time trend
Simple randomization	0.046	0.050
Adaptive randomization	0.049	0.205

2. Response and covariate adaptive randomization

Assume that the clinical trial consists of two treatments and n patients. We will first assume that n is pre-specified at the start of the trial and will later indicate how this restriction can be removed. Patients are registered for the trial one at a time. For the i th patient registered let y_i denote the outcome, c_i be a 0–1 indicator of the treatment assigned (0 for control, 1 for new treatment) and x_i be a vector of covariates measured before treatment assignment.

Let H_i denote the data available when the i th patient registers for the study and is ready for treatment assignment. This data consists of the measured covariates x_1, \dots, x_i for the current patient and all previously registered patients and the treatment assignments c_1, \dots, c_{i-1} for all previously registered patients. H_i also consists of the outcomes for some if not all of the previously registered patients y_1, \dots, y_{i-1} . Since outcome takes time to observe, H_i may include outcomes for only some of the previous patients. For simplicity of notation here, we will assume that there are no delays in observing response but our results below do not depend on this assumption.

We assume that the i th patient is assigned treatment 1 with probability $g(H_i)$ where g is a pre-specified function that defines the adaptive assignment mechanism. For equal non-adaptive randomization, $g(H_i) = 0.5$ for all $i = 1, \dots, n$. For “adaptive stratification” methods such as that developed by Pocock and Simon (1975), $g(H_i)$ depends on the covariate vectors x_1, x_2, \dots, x_i and the treatment assignments c_1, c_2, \dots, c_{i-1} but not on the outcomes y_1, y_2, \dots, y_{i-1} . For response adaptive methods, the randomization weights vary based on the measured covariates for the current patient and all previously registered patients, the treatment assignments for previously measured patients, and the outcomes for all previously registered patients which are available at the time of entry of the i th patient. The requirement that the g function be specified in advance is consistent with current regulatory guidelines.

At the time of analysis we assume that we have data D consisting of outcomes, measured covariates, and treatment assignments for all n patients. We denote this by $D(\underline{y}, \underline{x}, \underline{c})$. From this data we can compute a test statistic $T(D)$. For example, this may be a standardized difference in average outcome between the two groups, perhaps adjusted for measured covariates. For survival data, $T(D)$ might be a Wald statistic from a proportional hazards model incorporating treatment and covariates.

Table 1 shows the result of a simple simulation of two arm clinical trials with data for $n = 50$ patients. The test used for comparing the groups is based on the Mann–Whitney statistic. The observed difference was considered statistically significant if the large sample normal approximation was significant at a one-sided 5% level. For the first row, treatment assignment was based on simple equally weighted non-adaptive randomization. For the first column of the table, the outcomes y_1, y_2, \dots, y_n are independent and normally distributed with mean zero and variance 1; there are no measured covariates and no treatment effect. In this case, the type I error, estimated from 10,000 replicated trials, approximates the nominal 5% significance level used for the tests. The last column shows results when there is an unknown time trend. That is, y_i was normally distributed with mean $10i/n$ and variance 1. Again there were no measured covariates and no treatment effect. With or without time trend, using equally weighted non-adaptive randomization, the proportion of the 10,000 simulation replications in which the null hypothesis was rejected is approximately 0.05, and the small discrepancy is within the limitations of the number of replications and the accuracy of the large sample approximation to the Mann–Whitney statistic for clinical trials of only 50 patients.

The second row of Table 1 shows results for similar trials using a response adaptive randomization method. We assume that there is no delay in observing responses so H_i consists of outcomes and treatment assignments for patients $1, 2, \dots, i - 1$. The first 10 patients are assigned treatment using simple equally weighted randomization. For subsequent patients the randomization weight $g(H_i)$ is the standardized Mann–Whitney statistic for comparing outcomes for the two treatments using data for patients $1, 2, \dots, i - 1$. This standardized statistic equals the sum of the ranks for outcomes on treatment $c = 1$ minus $n_1(n_1 + 1)$ divided by $n_1 n_0$ where n_1 and n_0 denote the number of the first $i - 1$ patients who received treatments 1 and 0 respectively. This standardized statistic takes values in the range 0–1.

The inflation in the type I error shown in Table 1 for adaptive randomization with time trend results from the fact that the null hypothesis tested by the Mann–Whitney test is in fact false in that case. The null hypothesis is that the n_1 observations in treatment arm 1 and the n_2 observations in treatment arm 0 can be regarded as the selection of a partition of the data into sets of size n_1 and n_2 in which all such partitions are equally likely. This is not true for adaptive randomization. The inflation of the type I error can be understood intuitively from the following thought experiment involving a more extreme form of adaptive randomization. Suppose that the outcomes for the first 10 patients are sampled independently from $N(0, 1)$ and that these patients are randomly partitioned into 5 patients who receive treatment 1 and 5 who receive treatment 0. Let all of the remaining 40 patients be assigned to receive the treatment with the larger sum of ranks in the initial 10 patients. Suppose also that the outcomes for the 40 remaining patients are independent $N(\lambda, 1)$ observations. If λ is large enough, then the outcomes for all 40 additional patients will exceed the greatest outcome for the initial 10 patients. Suppose for example that

Table 2
Type I error for adaptive randomization test.

	No time trend	Time trend
Adaptive randomization	0.049	0.050

treatment group 1 has the smaller rank sum r_1 and that treatment 0 is thus assigned to the remaining 40 patients. Under the null hypothesis that the 5 observations in treatment group 1 are randomly selected from the 50 total observations, the probability that the rank sum for that group is no greater than r_1 is $M(r_1; 5, 10) / \binom{50}{5}$ where the numerator represents the number of combinations in which 5 observations can be randomly selected from the first 10 to result in a rank sum no greater than r_1 . The numerator is strictly less than $\binom{10}{5}$ and hence the probability is less than $\binom{10}{5} / \binom{50}{5}$ which is approximately 0.00012. Hence, the two-sided type I error will approach 1.0 for large values of λ . In order to avoid this inflation in type I error, an alternative analysis is needed that takes account of the adaptive randomization utilized in the clinical trials.

3. A significance test based on the randomization distribution

Let $dF_{\underline{c}|\underline{z}}$ denote the distribution of the sequence of treatment assignments $\underline{c} = (c_1, \dots, c_n)$ conditional on $\underline{z} = ((x_1, y_1), \dots, (x_n, y_n))$, the sequence of covariate vectors $\underline{x} = (x_1, \dots, x_n)$ and outcomes $\underline{y} = (y_1, \dots, y_n)$. One can sample from $dF_{\underline{c}|\underline{z}}$ under the null hypothesis by holding fixed the sequence of covariate vectors and outcomes for the patients in the clinical trial and re-randomizing all of the patients using the probabilistic treatment assignment mechanism determined by the adaptive algorithm. The sequence of treatment assignments sampled will in general depend on the sequence of covariate vectors and outcomes and these are kept fixed.

Let $dF_{T(\underline{z})}$ denote the distribution of the test statistic T induced when the vector of treatment assignments is drawn from $dF_{\underline{c}|\underline{z}}$. This induced distribution can be used as a null distribution for the test statistic computed from the data using the treatment assignments actually used in the clinical trial. For a one-sided significance test of level α of the null hypothesis against the alternative that treatment 1 is superior, we use as critical value for the test statistic the $100(1 - \alpha)$ th percentile of $dF_{T(\underline{z})}$, i.e. $F_{T(\underline{z})}^{-1}(1 - \alpha)$.

Table 2 shows the results of a simulation of the type I error obtained using this randomization test based on the adaptive assignment rule and test statistic described for Table 1. These results are based on 5000 clinical trials simulated under the two conditions described for Table 1. For each clinical trial, the null distribution of the test statistic was approximated based on repeating the adaptive treatment assignment 500 times. As can be seen, the adaptive randomization test has the correct type I error, even when there are time trends in the outcomes.

The following theorem provides conditions on the stochastic process generating the sequence $\underline{z} = ((x_1, y_1), \dots, (x_n, y_n))$ that ensures that the type I error is controlled at level α in hypothetical replications of the clinical trial with different covariate and outcome vectors. The controlled size of the rejection region for the randomization test conditional on \underline{z} does not automatically ensure control of the type I error under all conditions. If a test conditional on the full vector of outcomes \underline{z} has size $\leq \alpha$, then the type I error (i.e. expected size with regard to distribution of \underline{z}) will be $\leq \alpha$. The subtlety here comes from the fact that under arbitrary dependence structures, the randomization mechanism used for determining the critical value for the randomization test does not necessarily correspond to the conditional distribution of $\underline{c}|\underline{z}$. If z_j is dependent on c_i for $j > i$, then this induces a dependence in the other direction and information about z_j gives us information about c_i . The adaptive randomization test, however, determines c_j using no information about z_j for any $j > i$. Consequently, in order for the randomization distribution to match the true conditional distribution, we need (conditional) independence between c_i and z_j for $j > i$.

Theorem 1. Let $\underline{z} = ((x_1, y_1), \dots, (x_n, y_n))$ be a sequence of pairs of covariate vectors and outcomes and let $dF_{\underline{z}, \underline{c}}$ denote the joint distribution of \underline{z} and the vector of treatment assignments \underline{c} . Let $T(\underline{z}, \underline{c})$ denote the value of the test statistic computed on the data and $F_{T(\underline{z})}$ denote the distribution function of the null distribution of the test statistic T induced by the randomization process conditional on \underline{z} . For each $i \in \{1, \dots, n\}$, we assume that conditional on $((x_1, y_1), \dots, (x_{i-1}, y_{i-1}))$, (x_i, y_i) is independent of (c_1, \dots, c_{i-1}) . Then under the null hypothesis,

$$\Pr_{\underline{z}, \underline{c}} \left[T(\underline{z}, \underline{c}) \leq F_{T(\underline{z})}^{-1}(1 - \alpha) \right] \geq 1 - \alpha. \tag{1}$$

Proof.

$$\Pr_{\underline{z}, \underline{c}} \left[T(\underline{z}, \underline{c}) \leq F_{T(\underline{z})}^{-1}(1 - \alpha) \right] = \int_{\underline{z}, \underline{c}} I \left\{ T(\underline{z}, \underline{c}) \leq F_{T(\underline{z})}^{-1}(1 - \alpha) \right\} dF_{\underline{z}, \underline{c}} \tag{2}$$

where $I\{\}$ denotes the indicator function. The joint distribution can be written

$$\begin{aligned} dF_{\underline{z}, \underline{c}} &= dF_{z_1, c_1} dF_{z_2, c_2 | z_1, c_1} dF_{z_3, c_3 | z_1, z_2, c_1, c_2} \cdots dF_{z_n, c_n | z_1, \dots, z_{n-1}, c_1, \dots, c_{n-1}} \\ &= dF_{z_1} dF_{c_1 | z_1} dF_{z_2 | z_1, c_1} dF_{c_2 | z_1, z_2, c_1} \cdots dF_{z_n | z_1, \dots, z_{n-1}, c_1, \dots, c_{n-1}} dF_{c_n | z_1, \dots, z_n, c_1, \dots, c_{n-1}}. \end{aligned}$$

Conditional on z_1, \dots, z_{i-1} , the random vector z_i is assumed statistically independent of the treatment assignments c_1, \dots, c_{i-1} . Consequently,

$$\begin{aligned} dF_{\underline{z}, \underline{c}} &= dF_{z_1} dF_{c_1 | z_1} dF_{z_2 | z_1} dF_{c_2 | z_1, z_2, c_1} \cdots dF_{z_n | z_1, \dots, z_{n-1}} dF_{c_n | z_1, \dots, z_n, c_1, \dots, c_{n-1}} \\ &= dF_{\underline{z}} dF_{c_1 | z_1} dF_{c_2 | z_1, z_2, c_1} \cdots dF_{c_n | z_1, \dots, z_n, c_1, \dots, c_{n-1}}. \end{aligned}$$

The null hypothesis is $dF_{y_i | H_i, c_i} = dF_{y_i | H_i}$ for each i where H_i is as previously defined. This implies that $dF_{c_i | H_i, y_i} = dF_{c_i | H_i}$. Consequently,

$$\begin{aligned} dF_{\underline{z}, \underline{c}} &= dF_{\underline{z}} dF_{c_1 | x_1} dF_{c_2 | z_1, x_2, c_1} \cdots dF_{c_n | z_1, \dots, z_{n-1}, x_n, c_1, \dots, c_{n-1}} \\ &= dF_{\underline{z}} \prod_{i=1}^n dF_{c_i | H_i}. \end{aligned} \tag{3}$$

Using (3) in (2), we obtain

$$\Pr_{\underline{z}, \underline{c}} \left[T(\underline{z}, \underline{c}) \leq F_{T(\underline{z})}^{-1}(1 - \alpha) \right] = \int_{\underline{z}} dF_{\underline{z}} \int_{\underline{c}} I \left\{ T(\underline{z}, \underline{c}) \leq F_{T(\underline{z})}^{-1}(1 - \alpha) \right\} \prod_{i=1}^n dF_{c_i | H_i}. \tag{4}$$

By construction of the randomization test, however,

$$\int_{\underline{c}} I \left\{ T(\underline{z}, \underline{c}) \leq F_{T(\underline{z})}^{-1}(1 - \alpha) \right\} \prod_{i=1}^n dF_{c_i | H_i} \geq 1 - \alpha$$

and hence

$$\Pr_{\underline{z}, \underline{c}} \left[T(\underline{z}, \underline{c}) \leq F_{T(\underline{z})}^{-1}(1 - \alpha) \right] \geq \int_{\underline{z}} (1 - \alpha) dF_{\underline{z}} = (1 - \alpha).$$

Consequently, the type I error is controlled at level α where α is the significance level used to define the one-sided rejection region for the conditional randomization tests. \square

The conditions of the theorem are quite general. This generality pertains to the type of adaptive assignment, the test statistic and to the stochastic mechanism for generating the sequence of covariate vectors and outcomes. The mechanisms include time trends of arbitrary shapes. The assumption that conditional on $((x_1, y_1), \dots, (x_{i-1}, y_{i-1})), (x_i, y_i)$ is independent of (c_1, \dots, c_{i-1}) is not completely innocuous but will often be satisfied. Treatment assignments are not publicized even in un-blinded clinical trials. We performed a simulation to evaluate the extent of bias obtained with using the randomization analysis when this assumption is violated. The outcome y_i was simulated as $p_i + \varepsilon$ where ε is $N(0, 1)$ and p_i is the probability of assigning treatment 1 to the i th patient. This is a clear case of statistical dependence in violation of the assumption of the theorem, yet in 1000 simulations using the randomization test, the type I error was 0.064. This suggests that although the type I error is not bounded in this situation by the level of the randomization test, the magnitude of bias is not large.

4. Statistical power

We performed simulations to evaluate the statistical power of the randomization test. Outcome for the i th patient was normal with mean $\delta c_i + \beta (i/n)$ and variance 1. c_i is the binary treatment indicator, δ is the treatment effect and β is the slope of the time trend in prognosis among patients. Results are shown for total sample sizes of 50 and 100 patients. Treatment effect δ was selected to have power 0.90 for non-adaptive randomization using a Mann–Whitney test with one-sided level of 0.05. The δ values used depended on n and β . Power values shown in Table 3 are for adaptive randomization using the randomization test for analysis. The initial 20% of patients were randomized with equal probability to each treatment. Subsequently, the adaptive randomization was based on the standardized Mann–Whitney statistic previously described, but with a maximum (minimum) probability of assignment to treatment group 1 of “cap” (1-cap).

Table 3 indicates that the statistical power associated with using adaptive treatment assignment is lower than that with non-adaptive randomization (0.90). Without time trends, the 14% point reduction in power for the very small sample size of 50 total patients is greater than the 5% point reduction for a clinical trial of 100 patients. For $n = 50$, the reduction in power was reduced by capping the degree of adaptiveness of the randomization. For $n = 100$, such capping was less necessary and had less effect. Similar effects were seen with time trends in the prognosis of patients. The results for power shown in Table 3 are based on a particular method for adaptively modifying randomization weights and a particular test statistic. Other choices of weights and other test statistics might perform very differently.

Table 3
Power for adaptive randomization.

Total sample size (n)	Time trend (β)	Power cap = 1	Power cap = 0.67
50	0	0.76	0.82
50	1	0.73	0.80
100	0	0.85	0.85
100	1	0.83	0.84

5. Discussion

The randomization test we have described can be used with categorical, right-censored or continuous outcome data and with any test statistic. It can also be used with a wide variety of adaptive assignment mechanisms. The randomized play the winner rule developed by [Wei and Durham \(1978\)](#) was used in the controversial ECMO (extracorporeal membrane oxygenation) clinical trial for newborns with respiratory failure ([Ware and Epstein, 1985](#)). Wei and Durham developed a randomization test for that particular design with binary endpoint and it was discussed by [Begg \(1990\)](#).

To the best of our knowledge, this paper provides the first proof of the finite sample conditions under which the type I error is preserved for response adaptive treatment assignment mechanisms. [Jennison and Turnbull \(1999\)](#) and [Karrison et al. \(2003\)](#) described group sequential adaptive treatment designs in which randomization weights are constant for sequentially accrued blocks of patients. They demonstrated the preservation of type I error asymptotically for their treatment assignment rules, but that demonstration requires that the sample size of each block approaches infinity because the arguments are conditional on the weights used for the blocks.

In our derivation we assumed that the total sample size n was fixed in advance. Using the spending function approach of [Lan and DeMets \(1983\)](#) however, one can define a sequence of nominal significance thresholds $\alpha_1, \dots, \alpha_K$ for use at K analysis times with pre-specified sample sizes n_1, n_2, \dots, n_K . Using the randomization distribution of $dF_{\mathcal{L}|\mathcal{Z}}$ one can define critical values $T_k^*(z_1, \dots, z_{n_k})$ for size α_k rejection regions at these analysis times. In performing the k th analysis of a particular clinical trial, only the data $\{c_i, z_i, i = 1, \dots, n_k\}$ is needed in order to perform the randomization test and determine whether the test statistic exceeds the critical value $T_k^*(z_1, \dots, z_{n_k})$ and the proof of the theorem can be generalized to accommodate the new rejection region. With right-censored data, further modification to the derivation would be necessary.

For the calculations of our examples, we have approximated the significance levels of the randomization tests by randomly selecting 500 re-randomizations of treatment assignments. Efficient exact network algorithms and large sample approximations for some designs with binary outcomes have been studied by [Mehta et al. \(1988\)](#).

The randomization test approach can also be used for the analysis of clinical trials in which probabilistic adaptive stratification methods have been used for treatment assignment. Although no randomization test is possible for deterministic assignment methods like “minimization” ([Taves, 1974](#)), the Pocock–Simon adaptive stratification method ([Pocock and Simon, 1975](#)) is based on a biased coin randomization to balance the treatment groups marginally with regard to a large number of covariates. Although the method has been widely used, controversies occasionally arise regarding how the trials should be analyzed and what the effect of adaptive stratification is on type I error. The proof provided here may help to put such concerns to rest.

In this paper we have shown that if one ignores the adaptive assignment mechanisms in the analysis of response adaptive clinical trials, the type I error can be enormously increased by simple time trends in unmeasured characteristics of the patients. Adjustment for measured covariates may reduce this bias somewhat. Analyzing the data using a randomization test generated by the adaptive assignment mechanism enables the type I error to be controlled at the level of the randomization test under the conditions defined in [Theorem 1](#). These conditions include the existence of complex time trends in the data but the conditions are not completely general. Although this provides an improved statistical underpinning for the use of response adaptive treatment assignment, there are many other aspects that need to be considered in deciding whether to use such an approach such as effect on statistical power or on the number of patients treated with the inferior treatment in cases where the null hypothesis is false. We have provided some simulation results that indicate that the effect of response adaptive assignment and time trends on power can be substantial, particularly for very small clinical trials. Effect on power depends, however, on the nature of the outcome endpoint, the test statistic and how the randomization weights are modified by outcomes. Future evaluations of outcome adaptive randomization should, however, be based on the use of appropriate randomization tests that ensure that the target type I error is controlled at the desired level.

References

- Armitage, P., 1985. The search for optimality in clinical trials. *International Statistical Review* 53, 14–24.
- Begg, C.B., 1990. On inferences from Wei’s biased coin design for clinical trials. *Biometrika* 77 (3), 467–484.
- FDA, 2010. Draft guidance for industry: adaptive design clinical trials for drugs and biologics. Rockville, MD.
- Hoel, D.G., Sobel, M., Weiss, G.H., 1975. Comparison of sampling methods for choosing the best binomial population with delayed observations. *Journal of Statistical Computation and Simulation* 3 (4), 299–313.
- Jennison, C., Turnbull, B.W., 1999. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall.
- Karrison, T.G., Huo, D., Chappell, R., 2003. A group sequential, response-adaptive design for randomized clinical trials. *Controlled Clinical Trial* 24, 506–522.

- Lan, K.K.G., DeMets, D.L., 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70, 659–663.
- Mehta, C.R., Patel, N.R., Wei, L.J., 1988. Constructing exact significance tests with restricted randomization rules. *Biometrika* 75 (2), 295–302.
- Pocock, S.J., Simon, R., 1975. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 31, 103–115.
- Rosenberger, W.F., Lachin, J.M., 1993. The use of response-adaptive designs in clinical trials. *Controlled Clinical Trial* 14 (6), 471–484.
- Simon, R., 1977. Adaptive treatment assignment methods and clinical trials. *Biometrics* 33, 743–749.
- Taves, D.R., 1974. Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics* 15, 443–453.
- Ware, J.H., Epstein, M.F., 1985. Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics* 76, 849–851.
- Wei, L.J., Durham, S., 1978. The randomized play the winner rule in medical trials. *Journal of American Statistical Association* 73, 840–843.
- Zelen, M., 1969. Play the winner rule and the controlled clinical trial. *Journal of American Statistical Association* 64, 131–146.