Randomization-Based Inference: The Forgotten Component of the Randomized Clinical Trial

William F. Rosenberger

University Professor and Chairman
Department of Statistics
George Mason University

May 1, 2018

Why do we randomize?

Cornfield, "Principles of Research," *Journal of Chronic Diseases*, 1959.

> *1. It controls the probability that the treated and control groups differ more than a calculable amount in their exposure to disease, in immune history, or with respect to any other variable, known or unknown to the experimenter, that may have a bearing on the outcome of the trial. This calculable difference tends to zero as the size of the two groups increase.*
>
> *2. It makes possible, at the end of the trial, the answer to the question "In how many experiments could a difference of this magnitude have arisen by chance alone if the treatment truly has no effect?" It may seem mysterious that a mathematician could actually predict the course of future experiments. All you have to do is compute what would happen if a given set of numbers were randomly allocated in all possible ways between the two groups. Randomization allows this.*

Why do we randomize?

Point 1: The first property of randomization is that it promotes comparability among the study groups. Such comparability can only be attempted in observational studies by adjusting for or matching on *known* covariates, with no guarantee or assurance, even asymptotically, of control for other covariates. Randomization, however, extends a high probability of comparability with respect to unknown important covariates as well.

Despite the fact that consistent, replicated observational studies can also lead us to determine causality, there may always be questions as to whether we have controlled for all factors relating to incidence and prognosis of a disease. The randomized clinical trial allows this control, and hence represents the highest standard of evidence among biomedical studies.

Why do we randomize?

Point 2: The act of randomization provides a probabilistic basis for an inference from the observed results when considered in reference to all possible results. This randomization approach to inference is very different from the usual testing of unknown parameters arising from an independent and identically distributed sample from a known distribution. This is not taught in many biostatistical/statistical departments, and is the focus of this talk.

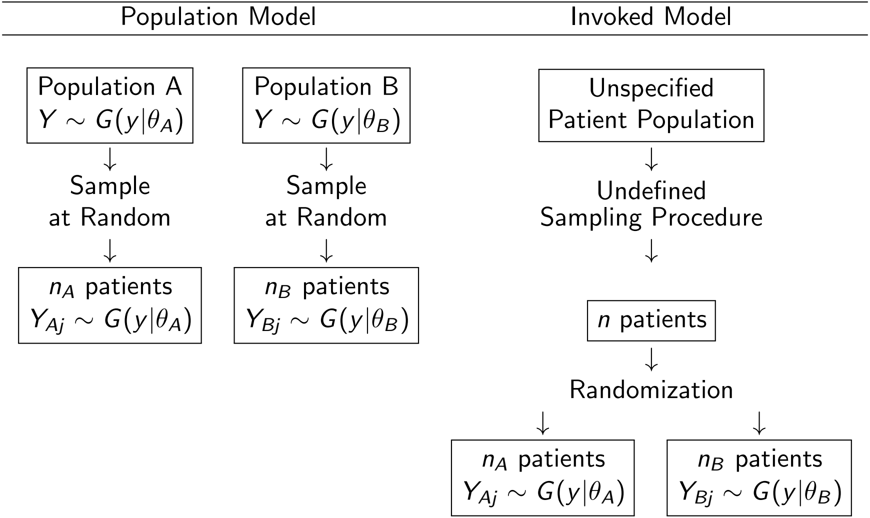
Why do we randomize?

I would add a Point 3: It provides a measure of unpredictability to the treatment assignment process. This unpredictability protects from certain biases that may enter the trial intentionally, unintentionally, consciously, or subconsciously.

Randomization as a Basis for Inference

- Cornfield's second great property of randomization is that it provides a basis for inference that is assumption free and relies only on the way in which the subjects were randomized.
- The early clinical trialists were aware of the importance of randomization-based inference, but had limited computer resources to implement it. Nowadays, we can run a randomization test (or "re-randomization test") in seconds, just by modifying the program used to generate the initial sequence.
- Unfortunately, students are not generally taught randomization tests, or even told that the usual population model does not apply to clinical trials.
- The absence of randomization-based inference from modern analyses is the principal reason that randomization merits only a sentence or two in medical journals.

The Population Model



Population Model

Invoked Model

Population A
$Y \sim G(y|\theta_A)$

Sample
at Random

n_A patients
$Y_{Aj} \sim G(y|\theta_A)$

Population B
$Y \sim G(y|\theta_B)$

Sample
at Random

n_B patients
$Y_{Bj} \sim G(y|\theta_B)$

Unspecified
Patient Population

Undefined
Sampling Procedure

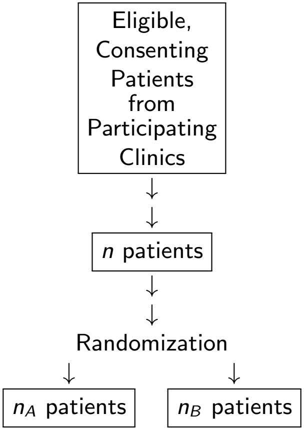
n patients

Randomization

n_A patients
$Y_{Aj} \sim G(y|\theta_A)$

n_B patients
$Y_{Bj} \sim G(y|\theta_B)$

The Randomization Model



Eligible, Consenting Patients from Participating Clinics

n patients

Randomization

n_A patients

n_B patients

The Randomization Model

As stated by Lachin (1988, p. 296):

> *The invocation of a population model for the analysis of a clinical trial becomes a matter of faith that is based upon assumptions that are inherently untestable.*

Fortunately, the use of randomization provides the basis for an assumption-free statistical test of the equality of the treatments among the n patients actually enrolled and studied. These are known as randomization tests.

Randomization Tests

The null hypothesis of a randomization test is that the assignment of treatment A versus B had no effect on the responses of the n patients randomized in the study. This *randomization null hypothesis* is very different from a null hypothesis under a population model, which is typically based on the equality of parameters from known distributions.

Randomization Tests

The essential feature of a randomization test is that, under the randomization null hypothesis, the set of observed responses is assumed to be a set of deterministic values that are unaffected by treatment. That is, under the null, each patient's observed response is what would have been observed regardless of whether treatment A or B had been assigned. Then the observed difference between the treatment groups depends only on the way in which the n patients were randomized.

Randomization Tests

One then selects an appropriate measure of the treatment group difference, or the treatment effect, which is used as the test statistic. The test statistic is then computed for all possible permutations of the randomization sequence. One then sums the probabilities of those randomization sequences whose test statistic values are at least as extreme as what was observed. This total is then the probability of obtaining a result at least as extreme as the one that was observed, which, by definition, is precisely the *p*-value of the test.

Randomization Tests

The key components of the validity of randomization-based inference is the randomization null hypothesis and the probability distribution induced by the randomization procedure itself. Standard population-based ideas such as the likelihood are replaced with the *reference set* induced by the randomization procedure: all possible sequences and their associate probabilities. Unlike in permutation testing, there is no assumption that each sequence is equiprobable, and, in fact, we must use the actual probabilities for the test to be valid.

Nonequiprobable Randomization Procedures

Examples of nonequiprobable randomization procedures:

- Permuted block design filling blocks using the truncated binomial design;
- Permuted block designs where block sizes are randomly selected;
- Restricted randomization procedures such as Efron's biased coin design, Wei's urn design, Soares and Wu's big stick design;
- Response-adaptive randomization, where treatment assignment probabilities are selected according to previous patient's responses;
- Covariate-adaptive randomization, where treatment assignment probabilities are selected according to the degree of balance on certain known covariates.

Nonequiprobable Randomization Procedures

Table 1: *Four Treatment Assignments under Random Allocation Rule (RAR) and Truncated Binomial Design (TBD)*

Randomization Sequence	Data Permutation		Probability	
x_1, x_2, x_3, x_4	A	B	P_{RAR}	P_{TBD}
$AABB$	x_1, x_2	x_3, x_4	1/6	1/4
$ABAB$	x_1, x_3	x_2, x_4	1/6	1/8
$ABBA$	x_1, x_4	x_2, x_3	1/6	1/8
$BAAB$	x_2, x_3	x_1, x_4	1/6	1/8
$BABA$	x_2, x_4	x_1, x_3	1/6	1/8
$BBAA$	x_3, x_4	x_1, x_2	1/6	1/4

Nonequiprobable Randomization Procedures

Efron's (1971) biased coin design: Gives a higher probability $p > 1/2$ of assigning the treatment that has the fewest assignments thus far. Let D_{j-1} be the difference in A and B numbers after $j-1$ patients have been randomized. Here

$$\begin{aligned} P(T_j = 1 | D_{j-1}) &= 1/2, \quad \text{if} \quad D_{j-1} = 0, \\ &= p, \quad \text{if} \quad D_{j-1} < 0, \\ &= 1-p, \quad \text{if} \quad D_{j-1} > 0. \end{aligned}$$

Efron suggested $p = 2/3$ might be a reasonable value (without justification).

Randomization Tests

Under the randomization null hypothesis treatments and responses are independent and all of these techniques can be analyzed using the same randomization-based inference techniques *with respect to the correct reference set*.

Note that, unlike in inference based on random sampling, I am completely unconcerned about the choice of test statistic, as long as it compares responses across treatment groups. I can use the difference of means or proportions, or a linear rank test. The advantage of a linear rank test is that it includes the Wilcoxon test, logrank test, and logrank test with censoring as special cases. I am also not concerned with the distribution of the chosen test, except with respect to the reference set. (*Try telling that to students!*)

What the Pioneers Thought

What did the Greats of Statistics think about the concept of randomization as a basis for inference? We begin with our hero:

Armitage (1954):

> *The customary test for an observed difference between two fatality rates is based on an enumeration of the probabilities, on the initial hypothesis that the two treatments do not differ in their effects, of all the various results that would occur if the trial were repeated indefinitely....*

What the Pioneers Thought

Efron noted in his 1971 paper that his biased coin design induces a nonequiprobable reference set, and so typical permutation tests do not apply:

> *The biased coin designs do not give the same conditional distribution as [complete randomization] and so (6.1) [the usual permutation testing formulation] does not apply directly. Theoretically we could redefine the rejection region of any permutation test to give level α with respect to the distribution [induced] under [the biased coin design],* **but this is hard work.**

What the Pioneers Thought

Cox (1982):

> *While the final analysis may not be based explicitly on the randomization distribution, it is necessary that there should be some broad correspondence with randomization theory.... It is now possible to test the null hypothesis as follows. Choose a suitable test statistic, such as the difference of means. Compare a suitable test statistic with the distribution induced by exact randomization probabilities.*

What the Pioneers Thought

The main concern by Efron and Wei was the actual computation of these tests for large samples. In the 1970s and 1980s computing resources would not allow exact enumeration or Monte Carlo approximations, and so the focus became asymptotic distributions. This is no longer relevant.

Armitage had another concern: how to incorporate a random stopping time into the reference set. This was a remarkable observation in 1954.

Cox's concern was that not all sequences in the reference set give very much information, and he suggested reducing the reference set to include only those sequences that contain close to the same numbers assigned to each treatment in the observed sequence. This led to a flurry of conditional limit theorems for randomization tests. We know how to do this pretty easily now using Monte Carlo procedures.

"The ABBA Example"

Table 2: *Unconditional and conditional reference sets for computation of the linear rank test from complete randomization. ABBA is the observed sequence.*

Unconditional ($\Omega_u = 16$)			Conditional ($\Omega_c = 6$)		
Sequence (l)	$\Pr(L = l)$	S_l	Sequence (l)	$\Pr(L = l)$	S_l
AAAA	1/16	0.0	*AABB*	1/6	−2.0
AAAB	1/16	−1.5	*ABAB*	1/6	0.0
AABA	1/16	−0.5	*ABBA*	1/6	1.0
AABB	1/16	−2.0	*BAAB*	1/6	−1.0
ABAA	1/16	1.5	*BABA*	1/6	0.0
ABAB	1/16	0.0	*BBAA*	1/6	2.0
ABBA	1/16	1.0			
ABBB	1/16	−0.5			
BAAA	1/16	0.5			
BAAB	1/16	−1.0			
BABA	1/16	0.0			
BABB	1/16	−1.5			
BBAA	1/16	2.0			
BBAB	1/16	0.5			
BBBA	1/16	1.5			
BBBB	1/16	0.0			

“The ABBA Example”

Suppose the observed sequence is $ABBA$ and the patient outcomes were $Y_1 = 3$, $Y_2 = 1$, $Y_3 = 4$, $Y_4 = 5$. Then the observed Wilcoxon rank-sum test statistics is 1.0. The one-sided p-values are $p_u = 4/16$ and $p_c = 2/6$.

“The ABBA Example”

Table 3: *Unconditional and conditional reference sets for computation of the linear rank test from Wei’s urn design.*

Unconditional ($\Omega_u = 16$)			Conditional ($\Omega_c = 6$)		
Sequence (l)	$\Pr(L = l)$	S_l	Sequence (l)	$\Pr(L = l)$	S_l
AAAA	0	0.0	*AABB*	0	-2.0
AAAB	0	-1.5	*ABAB*	1/4	0.0
AABA	0	-0.5	*ABBA*	1/4	1.0
AABB	0	-2.0	*BAAB*	1/4	-1.0
ABAA	1/12	1.5	*BABA*	1/4	0.0
ABAB	1/6	0.0	*BBAA*	0	2.0
ABBA	1/6	1.0			
ABBB	1/12	-0.5			
BAAA	1/12	0.5			
BAAB	1/6	-1.0			
BABA	1/6	0.0			
BABB	1/12	-1.5			
BBAA	0	2.0			
BBAB	0	0.5			
BBBA	0	1.5			
BBBB	0	0.0			

“The ABBA Example”

Now the sequences are not equiprobable. Here we compute $p_u = 1/12 + 1/6 + 0 + 0 = 1/4$ and $p_c = 1/4 + 0 = 1/4$.

Monte Carlo Randomization Test or “Re-Randomization Test”

For a set of observed responses $x_1, ..., x_n$ and the treatment assignments used in the trial $t_1, ..., t_n$, generated by a randomization procedure ϕ_j, we compute a test statistic, which can be based on any treatment effect difference, and call it $S_{obs.}$. Now we generate L randomization sequences using Monte Carlo simulation. For each of these sequences, a new test statistic, $S_l, l = 1, ..., L$, is computed from $x_1, ..., x_n$. The two-sided Monte Carlo p-value estimator is then defined as

$$\hat{p}_u = \frac{\sum_{l=1}^{L} I(|S_l| \geq |S_{obs.}|)}{L}. \tag{1}$$

For restricted randomization, the key component of this computation is that disparate probabilities of sequences will be depicted by the frequency of duplicate sequences sampled with replacement.

How Large Does L Have to Be?

Whether or not S_l is extreme is distributed as Bernoulli with underlying probability p_u, and hence $\hat{p}_u$ is unbiased with

$$MSE(\hat{p}_u) = \frac{p_u(1-p_u)}{L}.$$

Then establishing a bound $MSE(\hat{p}_u) < \epsilon$ implies that $L > 1/4\epsilon$. For $\epsilon = 0.0001$, we have $L > 2500$ (Zhang and Rosenberger (2011)).

The value of ϵ may not be small enough to estimate very small p-values accurately. Plamadeala and Rosenberger (2012) suggest finding L that ensures $P(|\hat{p}_u - p_u| \leq 0.1 p_u) = 0.99$, for instance. It follows that $L \approx (2.576/0.1)^2(1-p_u)/p_u$. Thus, to estimate a p–value as large as 0.04 with an error of 10% with 0.99 probability, the Monte Carlo sample size must be $L = 15,924$. If a smaller p-value is expected, L will be larger.

In any event, generating 20,000 randomization sequences takes only seconds.

Cox's Conditional Test

Cox recommended that we include in the reference set only those sequences that have the same number of treatment assignments to each arm as was observed. The naive approach would be to generate $M >> L$ sequences and keep only those that satisfy $N_A = n_{A,obs.}$, where M is large enough so that L is sufficient. This is prohibitively expensive computationally.

Cox's Conditional Test

Plamadeala and Rosenberger (2012) show how to do this for any restricted randomization procedure by generating sequences directly from the conditional reference set. Let $\phi_j(m_{j-1}) = P(T_j = 1 | N_A(j-1) = m_{j-1})$ be a randomization procedure. Then we generate sequences using the new randomization procedure

$$p_j = \begin{cases} \phi_j(m_{j-1}) \dfrac{P\left(N_A(n) = n_A | N_A(j) = m_j\right)}{P\left(N_A(n) = n_A | N_A(j-1) = m_{j-1}\right)}, & 1 < j \leq n, \\ \\ \dfrac{P\left(N_A(n) = n_A | T_j = 1\right)}{2P\left(N_A(n) = n_A\right)}, & j = 1. \end{cases}$$

In some cases, the conditional probabilities in this formula are hard to compute. This reduces the computational compexity down to the same level as for the unconditional reference set (i.e., generate L sequences).

Error Rates

Simulation of type I error rate and power can be done by replicating clinical trial outcome data M times, and determining the proportion of the M trials in which $p < \alpha$, where α is pre-specified. Outcome data can be generated from any homogeneous model under the null hypothesis, or under a different model for each treatment for the alternative hypothesis.

Error Rates

- By its construction, type I error rates are always preserved, as the p-value will be uniformly distributed over the randomization distribution under H_0, unless the study is biased resulting in incomparable groups with respect to a covariate.
- (Except for small samples due to discreteness.)
- Hence our friends in Vienna (Martin, Franz, Peter) should like randomization tests.
- What about power? One of the great aspects of randomization tests is that they tend to preserve power when the parametric assumptions of population-based tests are violated.
- However, power will depend a great deal on the particular randomization procedure used.

Error Rates

	Model (1)				Model (2)			
	Randomization		*t*-test		Randomization		*t*-test	
Procedure	Size	Power	Size	Power	Size	Power	Size	Power
CR	0.05	0.87	0.05	0.93	0.05	0.57	0.05	0.60
RAR	0.04	0.93	0.04	0.93	0.05	0.61	0.04	0.60
TBD	0.05	0.93	0.05	0.93	0.05	0.35	0.18	0.57
Smith ($\rho = 1$)	0.05	0.91	0.05	0.93	0.05	0.66	0.02	0.62
BCD	0.04	0.92	0.05	0.93	0.05	0.78	0.01	0.64
PBD	0.05	0.93	0.04	0.93	0.05	0.88	0.00	0.65
RBD	0.05	0.93	0.04	0.93	0.05	0.90	0.00	0.65
BSD	0.05	0.93	0.05	0.93	0.05	0.83	0.00	0.61

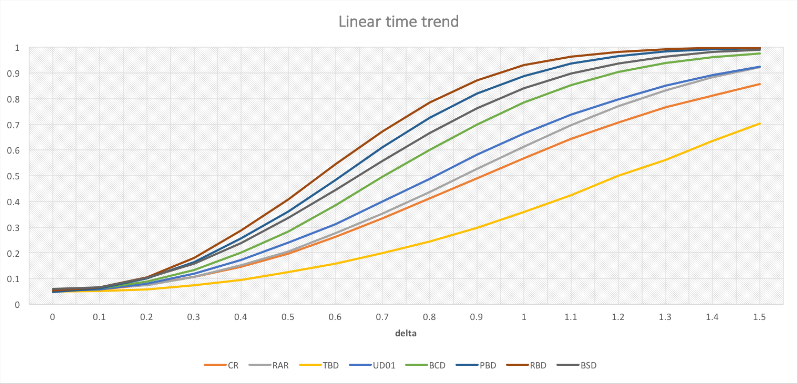


Figure 1: *Power of the randomization test (difference of means) under a linear time trend.* $H_0 : Y \sim N(0, 1) + (4i/n - 2)$*;*
H_A *: treatment A is* $N(\Delta, 1) + (4i/n - 2)$*. Each simulation is based on 10,000 tests, n=50, L=15,000.*

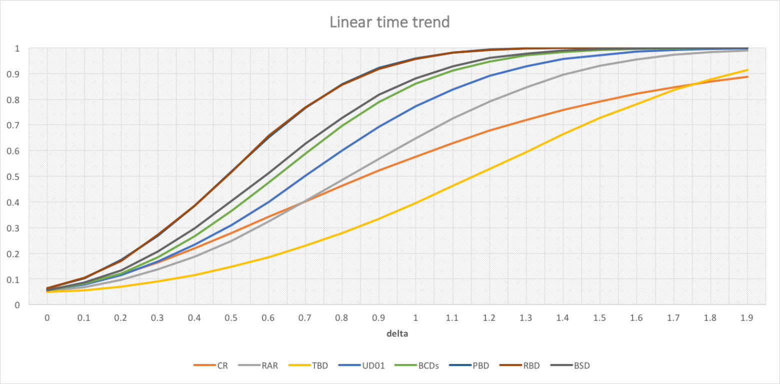


Figure 2: *Power of the randomization test (Wilcoxon test) under a linear time trend. $H_0 : Y \sim N(0, 1) + (4i/n - 2)$; H_A : treatment A is $N(\Delta, 1) + (4i/n - 2)$. Each simulation is based on 10,000 tests, n=50, L=15,000.*

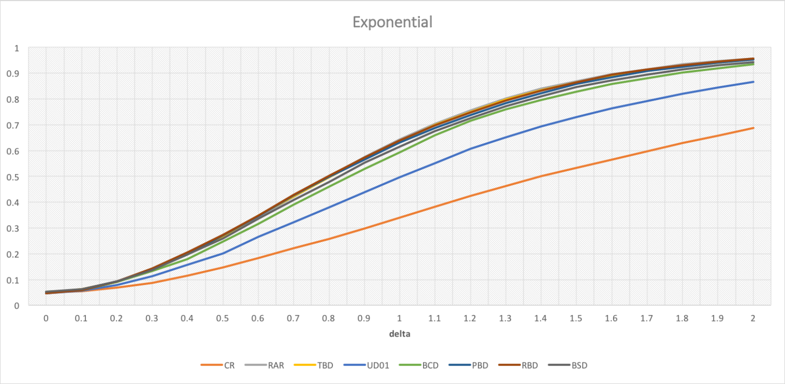


Figure 3: *Power of the randomization test (difference of means) under exponential response.* $H_0 : Y \sim exp(1)$; H_A : *treatment A is* $exp(\Delta + 1)$. *Each simulation is based on 10,000 tests, n=50, L=15,000.*

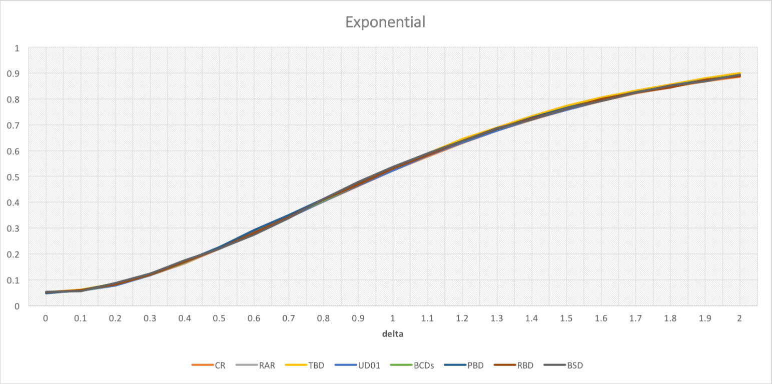


Figure 4: *Power of the randomization test (Wilcoxon test) under exponential response.* $H_0 : Y \sim exp(1)$; H_A : *treatment A is* $exp(\Delta + 1)$. *Each simulation is based on 10,000 tests, n=50, L=15,000.*

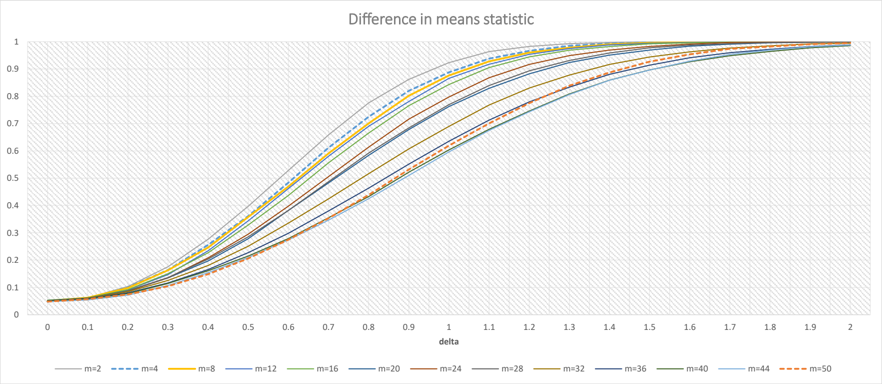


Figure 5: *Simulated power curves of the randomization tests under linear time trend for permuted block designs with different block size m. Each simulation is based on 10,000 tests, n=50, L=15,000.*

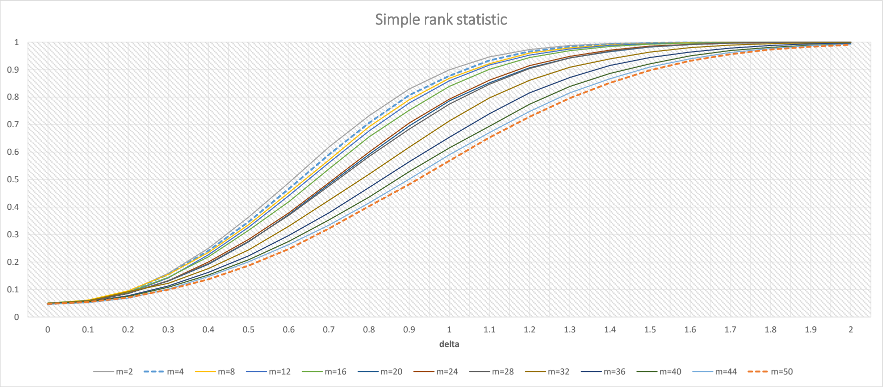


Figure 6: *Simulated power curves of the randomization tests under linear time trend for permuted block designs with different block size m. Each simulation is based on 10,000 tests, n=50, L=15,000.*

Summary

- Randomization has become a rote exercise that is nearly ignored in practice.
- Its basis for inference has been cited since the dawn of clinical trials as one of the key advantages of its use.
- Researchers in past decades have not been able to compute randomization tests due to computational limitations.
- The Monte Carlo formulation makes them available in seconds.
- Randomization-based inference can be used for virtually any primary outcome analysis encountered in clinical trials.
- Power in randomization-based inference is a property of the randomization distribution rather than the distribution of the test statistic.
- RandomizeR, created by Diane Uschner, is a new R package that computes randomization tests as well as properties of randomization procedures. It can revolutionize the design and analysis phase of the clinical trials.

Randomization Matters

Thank you!