Randomization Tests: The Forgotten Component of the Randomized Clinical Trial

William F. Rosenberger

University Professor
Department of Statistics
George Mason University

Ingram Olkin Memorial NISS Symposium
September 15, 2020

Why do we randomize?

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

> *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.*

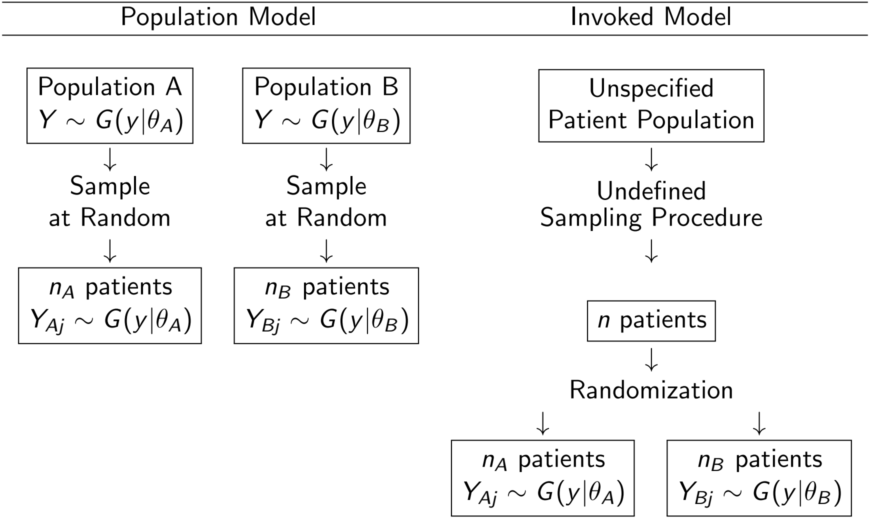
- At the initial session, we heard several speakers talk about the importance of preserving the “integrity of randomization” during the COVID crisis. What exactly does this mean, and why is it important?
- The CONSORT document indicates that all clinical trials must report the randomization procedure used and its impact on bias. It does not mention inference. There are still many authors who don’t know what a randomization procedure is, and try to satisfy this criterion by writing ”randomization was done by EXCEL,” or some other package.
- Most randomization procedures protect against subtle biases, especially in large samples (Rosenberger and Lachin, 2016), so what is the importance of specifying the randomization procedure? And surely discontinuing a randomized patient due to contracting COVID should not immediately interject bias by destroying the integrity of randomization. Patients are unblinded due to SAEs all the time.
- Are people really talking about unblinding rather than subverting randomization? I find that people mix these concepts up all the time.
- What is the point of specifying the randomization procedure and protecting that procedure against all enemies, foreign and domestic?

In the absence of using randomization as a basis for inference, there really is no point!

Randomization as a Basis for Inference

- 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

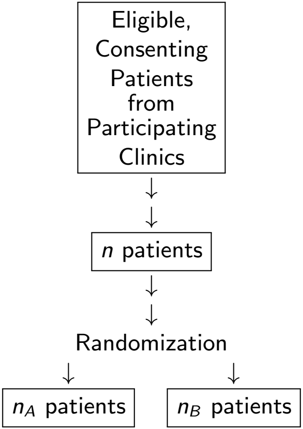
n patients

Randomization

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

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

The Randomization Model



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. Randomization tests are *not* permutation tests, where the data are presumed to arise from an exchangeable population.

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

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.

"The ABBA Example"

Table 2: *Reference sets for computation of the exact randomization test from complete randomization and Efron's biased coin design (BCD) with* $p = 2/3$*. The observed sequence is ABBA. The two-sided p-value for complete randomization is* 0.25 *and for the BCD is* 0.30*.*

Sequence (l)	$\Pr(L = l)$ complete	$\Pr(L = l)$ BCD	S_l
AAAA	0.0625	0.0185	145.0^*
AAAB	0.0625	0.0370	-33.3
AABA	0.0625	0.0370	46.7
AABB	0.0625	0.0741	10.0
ABAA	0.0625	0.0556	33.3
ABAB	0.0625	0.1111	0.0
ABBA	0.0625	0.1111	50.0^*
ABBB	0.0625	0.0556	46.7
BAAA	0.0625	0.0556	-46.7
BAAB	0.0625	0.1111	-50.0^*
BABA	0.0625	0.1111	0.0
BABB	0.0625	0.0556	-33.3
BBAA	0.0625	0.0741	-10.0
BBAB	0.0625	0.0370	-46.7
BBBA	0.0625	0.0370	33.3
BBBB	0.0625	0.0185	-145.0^*

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.

Regression Modelling

It may be desired to test a covariate-adjusted treatment effect from a regression model. While a regression model is usually developed under a population model, it is straightforward to apply a randomization analysis following the fit of a model. Conceptually the basic steps are to first fit a model to baseline covariates, other than treatment group, since the test is conducted under H_0.

Then the residuals from the model can be viewed as a set of fixed responses, regardless of which treatment is assigned. The model residuals can then be employed in lieu of the responses as the basis for computing a test statistic. This approach was first described by Gail, Tan, and Piantadosi (1988). They used the asymptotic distribution of the randomization test assuming treatment assignments are equiprobable.

Regression Modelling

A major advantage of the randomization analysis of the residuals is that the validity of the test in no way depends on the validity of the model assumptions used to fit the model. Thus, if a simple normal errors model is used as the basis for computing the residuals, then the validity of a t- or F-test between groups depends on the homoscedastic normal errors assumption. However, the randomization test comparing the randomly assigned groups in no way depends on this assumption. Thus the randomization test can be viewed as a robust test in situations where the regression model may be misspecified, unless the residuals computed are wrong due to extreme misspecification.

Regression Modelling

Parhat, Rosenberger, and Diao (2014) directly compute the randomization test by ranking the residuals and calculating the linear rank test, then re-randomizing them. They also apply the technique to time-to-event data using both the Cox proportional hazards model and the accelerated failure time model, by ranking the martingale residuals. They find that the randomization test preserves error rates better than tests based on the population model, when the underlying model is misspecified. They have a series of SAS macros that do this.

Group Sequential Monitoring

From a group sequential standpoint, we set up a series of boundary values $d_1, ..., d_K$ over K interim inspections, and we establish a spending function $\alpha(t_k), 0 < t_1 < t_2 < \cdots t_K = 1$ that "spends" a portion of the total α at each t_k, but preserves $\alpha(1) = \alpha$. At the kth inspection, we need to compute the conditional probability distribution of the test statistic S_k, conditional on $S_1 <= d_1, S_2 <= d_2, \ldots S_{k-1} <= d_{k-1}$.

The Monte Carlo randomization test procedure to do this is clear. We generate sequences and keep only those that satisfy the condition $S_1 <= d_1, S_2 <= d_2, \ldots S_{k-1} <= d_{k-1}$, and then determine d_k as a quantile of the randomization distribution of S_k such that $P(S_k > d_k)$ is the incremental alpha determined by $\alpha(t)$. The trial ends when the observed test statistic $S_{k,obs.} > d_k$. This is the approach of Plamadeala and Rosenberger (2012). The paper also addressed the tricky issue of how to define "information" when there is no concept of Fisher information in the randomization-based formulation.

The fully sequential approach has no analog, because *n* is then a random variable, and the reference set for the randomization test must include all variable length sequences that could have been realized. The only way to do this is to condition on *n*, but the stopping rule includes information on the treatment effect.

Interestingly, Armitage recognized this problem in his 1952 paper on sequential analysis!

Covariate-Adaptive Randomization

Covariate-adaptive randomization (sometimes known as *minimization*) randomizes patients with a certain covariate profile according to the degree of imbalance among treatment assignments for already randomized patients with the same profile. The idea is that one can incorporate far more covariates than is allowable using stratification. Taves (1974) proposed a non-randomized minimization design that does not allow randomization-based inference (except when there are ties). Pocock and Simon (1975) used Efron's biased coin design as a randomization procedure within marginal covariate profiles. In fact, Simon (1979) indicates how to use randomization-based inference following the procedure.

Covariate-Adaptive Randomization

It is possible, though cumbersome, to perform the appropriate randomization test generated by a nondeterministic adaptive stratification design. One assumes that the patient responses, covariate values, and sequence of patient arrivals are all fixed. One then simulates on a computer the assignment of treatments to patients using the [Pocock-Simon procedure] and the treatment assignment probabilities actually employed. Replication of the simulation generates the approximate null distribution of the test statistic adopted, and the significance level. One need not make the questionable assumption that the sequence of patient arrivals is random.

Preserving Operating Characteristics Under Heterogeneity

- One of the advantages of randomization tests is that they tend to preserve the type I error rate when the population model is misspecified.
- To investigate this in small samples, we simulated 10,000 test statistics ($n = 50$) under two models:
  1. Under H_0, $Z_1, ..., Z_n \sim$ i.i.d. $N(0, 1)$.
     Under H_1, treatment A has a mean shift of 1.
  2. Under H_0, $Z_1, ..., Z_n$ are subject to a drift over time, ranging linearly on the interval $(-2, 2]$ plus a $N(0, 1)$ random variable.
     Under H_1, treatment A has a means shift of 1.

	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

Recent papers

Statistics in Medicine has published four recent papers in the *Tutorial* section of the journal:

ROSENBERGER, W. F., USCHNER, D., and WANG, Y. (2019). Randomization: the forgotten component of the randomized clinical trial. Statist. Med. 38 1-30 (with discussion).

PROSCHAN, M. and DODD, L. E. (2019). Re-randomization tests in clinical trials. Statist. Med. 38 2292-2302.

WANG, Y., ROSENBERGER, W. F., and USCHNER, D. (2019). Randomization tests for multi-armed randomized clinical trials," Statist. Med. 39 494-509.

WANG, Y. and ROSENBERGER, W. F. (2020). Randomization-based interval estimation in randomized clinical trials. Statist. Med., 39, 2843–2854.

Some old papers:

All these ideas have their roots in papers by Anscombe and Armitage, and Kempthorne knew how to do everything, except actually compute them! Randomized clinical trials are perhaps the most perfectly suited experiments ever invented to take advantage of these simple, powerful, and type I error-preserving methods.

- Q: If the analysis of a clinical trial is based on a randomization model that does not in any way involve the notion of a population, how can results of the trial be generalized to determine the best care for future patients?
- A: Berger (2000) argues that the difficulty in generalizing to a target population is a weakness not of the randomization test, but of the study design. If it were suspected by investigators that patient experience in a particular clinical trial could not be generalized, there would be no reason to conduct the clinical trial in the first place. Thus we hope that the results of a randomized clinical trial will apply to the general population as well as to the patients in the trial. However, the study design only provides a formal assessment of the latter, not the former. By ensuring validity of the treatment comparison within the trial conducted, by limiting bias and ensuring strict adherence to the protocol, it is more likely that a generalization beyond the trial can be attained.

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.
- We have shown that randomization-based inference can be used for virtually any primary outcome analysis encountered in clinical trials.

Summary

- Randomization tests are valid even under heterogenity, preserving the type I error rate. This is not approximate, based on normal approximations, Z-tests, or multivariate normal assumptions. Because there are no Z-tests or assumptions.
- The only invalidity of a randomization test will be due to bias in the study.

The validity of the generalization of the experimental conclusions to a relevant population of interest relies on the design and proper conduct of the trial rather than on the accuracy of a statistical model of the population distribution. For example, the preservation of type I error rate may be viewed as a sufficient condition for valid generalization from the specific trial participants to a larger context. However, such a connotation cannot be invoked by the preservation of type I error rate alone; rather, a statistically valid conclusion may not be scientifically objective due to a biased experiment. It is sometimes also believed that tests involving random sampling from population distribution enables generalization to a broader context, whereas randomization tests, in which patient responses are regarded as arithmetic numbers, do not. Nonetheless, the population model only allows inference on the assumed population parameters or characteristics. –Wang, Rosenberger, Uschner, Stat. Med., *2019.*

Randomization Tests in the Time of COVID

COVID has upended clinical trials, and the most critical aspect is the introduction of heterogeneity: if a trial has an unplanned interruption, when it is restarted will the clinical trials population have the same characteristics? COVID diagnoses interfere with heart, lung, and blood functions. And recruitment may be limited to those who are at low risk. These aspects increase heterogeneity. Randomization-based inference can help because it retains operating characteristics under heterogeneity.

But the generalizability may be impacted.

Randomization Matters

Thank you!