Randomization in Small Clinical Trials

William F. Rosenberger

University Professor and Chairman, Department of Statistics, George Mason University

10 July 2015

Table of Contents

Table of Contents

Table of Contents

Introduction

- Rare diseases have low prevalence, less than 1 in 2000
- 6000 to 8000 rare diseases affect 30 million people in the EU
- About 50% are in children
- 80% genetic, 20% other causes

Source: Cornu, et al., 2013

Introduction

- Recruiting sufficient patients is difficult
- Appropriate subjects for clinical trials may be spread thinly within, or across countries
- Because many of the diseases are life-threatening and alternative therapies may not exist, there may be ethical imperatives to maximize the number of patients receiving the experimental therapy.
- Trials will, by necessity, be small, perhaps as few as 50 patients.

Some Design Considerations

- One common approach is to have $2:1$ allocation, so that $2/3$ of the patients receive active therapy.
- Using a standard sample size formula, this leads to an increase in total sample size of 13% in a trial that may already have recruitment problems.
- Another approach is to use a cross-over design, but it may be impossible to withdraw a patient from the experimental therapy if randomized to it in the first stage (e.g., chemotherapy) due to carryover effects, etc.
- Another approach is to have a two-stage trial, where only the first stage is counted and then everyone gets active therapy in the second stage.
- Sometimes the experimental therapy is not effective, or toxic, and so blindly increasing the number of patients assigned to it is a bad idea.

Impact on Randomization

- Some serious rare diseases are more homogeneous across doctors, clinics, and borders, potentially limiting the need for stratification. Also, a genetic marker may be present in every patient.
- The typical benefits of randomization: comparability on unknown underlying covariates, unpredictability to minimize bias, often rely on asymptotic properties that may not hold in small trials.
- While response-adaptive randomization may look especially attractive to assign more patients to the treatment which has performed the best in the trial, in small trials of 50 to 100, the benefit to patients would be so minimal to be almost meaningless.
- Standard homogeneous population models based on i.i.d. samples that are invoked for inference may not be appropriate for small clinical trials.

Purpose of this Presentation

- To investigate different randomization procedures for clinical trials, and see which procedures perform "best" for small clinical trials.
- Criteria of interest:
  1. Balance (efficiency)
  2. Comparability on unknown covariates
  3. Predictability
- To discuss randomization as a basis for inference and how randomization test can be used simply for many different types of inference.
- Note that results for $n = 50, 100$, while too small for many clinical trials, are still relevant because most multi-center trials are stratified, with small to moderate numbers of patients in each stratum.

Table of Contents

Complete Randomization

- Complete randomization is accomplished by tossing a fair coin, so the probability that patient j will receive treatment A is always $\phi_j = 1/2$.
- While complete randomization is completely unpredictable, there is the potential for large imbalances, both at the end of the trial and at middle points. In smaller trials, unusual sequences with large imbalances, or runs of a single treatment have a higher probability of occurring, perhaps not on average, but we are generating a single sequence.

Restricted Randomization

- For these reasons, restricted randomization is typically used, where ϕ_j is the probability that A is assigned to the jth patient, conditional on the first $j-1$ treatment assignments. Usually conditioning is done on the current number of patients assigned to treatment A, $N_A(j-1)$ (random), or on the difference in treatment numbers $D_{j-1} = 2N_A(j-1) - (j-1)$, to assign the next patient a higher probability of receiving the treatment that has had fewest assignments.
- To be a valid restricted randomization procedure, the unconditional probability of A must be $1/2$ for all n patients.

Forced Balance Procedures

- If we always knew n in advance, we could achieve exact balance while maintaining randomization using a forced balance procedure. In practice, we have a target sample size, but we must be prepared to randomize fewer or more patients than the target. So these procedures are difficult to apply directly in practice.
- Random allocation rule: An urn with $n/2$ A balls and $n/2$ B balls, n draws without replacement. Here

$$\phi_j = \frac{\frac{n}{2} - N_A(j-1)}{n-(j-1)}.$$

Forced Balance Procedures

- Truncated binomial design: Complete randomization is used until $n/2$ have been assigned to A or B, then the remainder is filled with the opposite treatment with probability 1. Here

$$\begin{aligned} \phi_j = &= 1/2, \quad \text{if} \quad \max\{N_A(j-1), N_B(j-1)\} < n/2, \\ &= 0, \quad \text{if} \quad N_A(j-1) = n/2, \\ &= 1, \quad \text{if} \quad N_B(j-1) = n/2, \end{aligned}$$

- Berger, et al.'s (2003) Maximal Procedure: Removes all sequences from the random allocation rule where $|D_j| > b$ for any $j = 1, ..., n$. The value b is called the *maximum imbalance parameter*. Each possible sequence remaining then has equal probability of being selected.

Blocked Procedures

- Permuted block design: Blocks of even size $2b$ are filled using either a random allocation rule or a truncated binomial design.
- The maximum imbalance is b and the only possibility of a terminal imbalance occurs if the last block is unfilled. Every block has at least one deterministic assignment.
- Random block design: Blocks of size $2, 4, 6, ..., 2K$ are randomly selected and equiprobable. The maximum imbalance is K, there is terminal balance if the last block is filled, and no assignments are made with probability 1.

Blocked Procedures

- Hadamard randomization: (Bailey and Nelson, 2003) A *Hadamard matrix* is a matrix of 1's and 0's in which two different rows have the same entries in half the columns and opposite entries in the other half.
- They suggest using a Hadamard matrix of size 12 with 1 row removed, and doubling the rows, switching 1's and 0's to create 22 possible randomization sequences which are equiprobable.
- The design should lessen the probability of imbalance in the final block, and has other nice properties in the ANOVA context.

Adaptive 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. Here

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

- Accelerated biased coin design: (Baldi Antognini and Giovagnoli, 2004) Adapts according to the magnitude of the imbalance of the biased coin design using a tuning parameter a:

$$\begin{aligned} \phi_{j} &= 1/2, && \text{if} \quad D_{j-1} = 0, \\ &= \frac{|D_{j-1}|^a}{|D_{j-1}|^a+1}, && \text{if} \quad D_{j-1} \leq -1, \\ &= \frac{1}{|D_{j-1}|^a+1}, && \text{if} \quad D_{j-1} \geq 1. \end{aligned}$$

Adaptive procedures

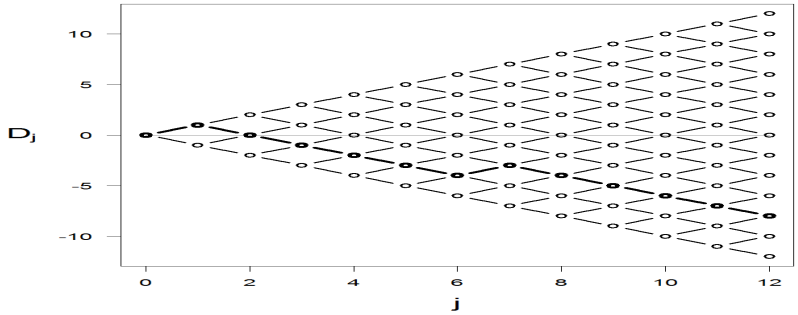
- Smith's (1984) design: Similar to ABCD, with tuning parameter ρ:

$$\phi_{j-1} = \frac{N_B(j)^\rho}{N_A(j)^\rho + N_B(j)^\rho}.$$

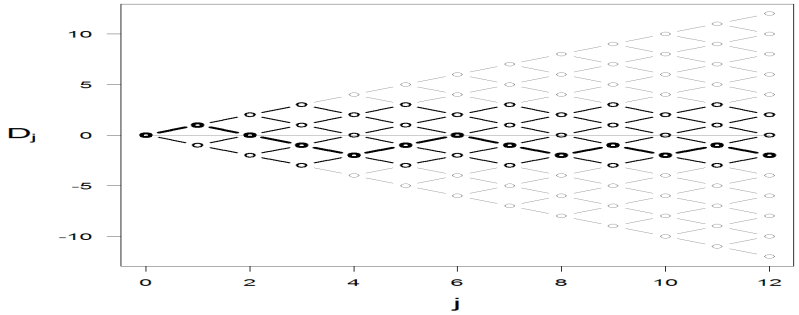
- If $\rho = 0$, this is complete randomization, and if $\rho = 1$ it is Wei's (1978) urn design.
- Designs with Imbalance Intolerance: By imposing a reflecting barrier, one can adapt these designs further.
- Big stick design (Soares and Wu, 1982): complete randomization with a reflecting barrier.
- BCD with Imbalance Intolerance (Chen, 1999): Efron's BCD with a reflecting barrier.
- Maximal Procedure: Random allocation rule with a reflecting barrier.

Table of Contents

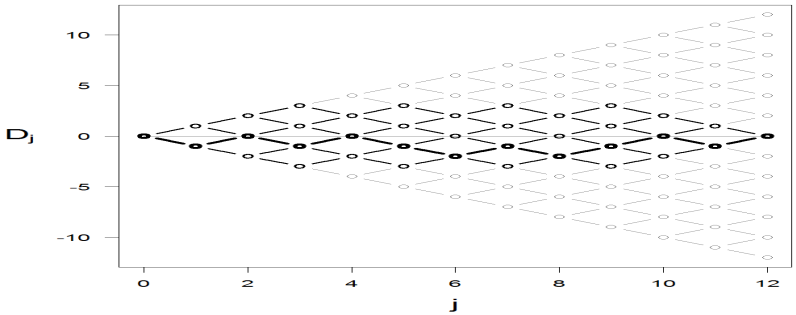
Comparison of Procedures



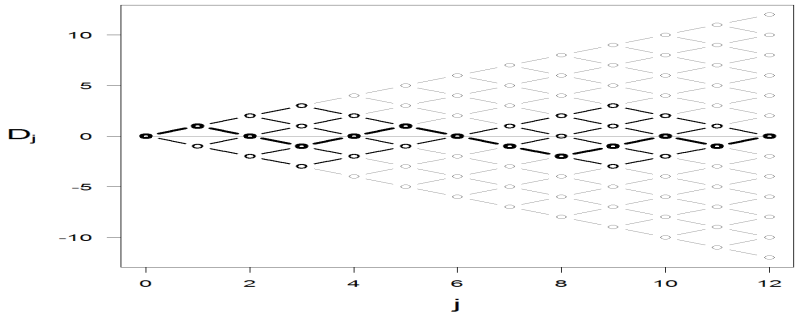
(a) *Efron's BCD* ($p = 2/3$)



(b) *Big stick design* ($b = 3$)



(c) *Maximal procedure* ($b = 3$)



(d) *Permuted block design* ($m = 6$)

Figure 1 : Possible paths (light bold) of four different randomization procedures, with a particular path outlined in heavy bold (thanks to Diane Uschner.)

Balancing Properties

Table 1 : *Simulated variance of D_n and expected maximum imbalance for eight different randomization procedures, $n = 50$, based on $100,000$ replications.*

Randomization procedure	$\text{Var}(D_n)$	$E\left(\max_{1\leq j\leq n}\lvert D_j\rvert\right)$
Complete	49.92	8.88
Smith($\rho = 1$)	16.58	5.83
Smith ($\rho = 5$)	4.69	3.77
$BCD(p = 2/3)$	4.36	4.28
Big Stick ($b = 3$)	2.66	3.00
$BCDII(p = 2/3, b = 3)$	1.70	2.94
$ABCD(a = 10)$	2.01	2.01
Hadamard	1.90	3.23
Random Block Design ($K = 3$)	0.75	2.34
Random Block Design ($K = 10$)	2.43	3.75

Comparability on Unknown Covariates

- The first great property of randomization is that it promotes comparability between treatment groups with respect to any unknown covariate, including those that may be correlated with patient outcomes. However, this is an asymptotic property. How well does this work in small trials?
- We examine three streams of covariates:
  (1) $Z_1, ..., Z_n$ are i.i.d. $N(0, 1)$.
  (2) $Z_1, ..., Z_n$ are subject to a drift over time, ranging linearly on the interval $(-2, 2]$ plus a $N(0, 1)$ random variable.
  (3) $Z_1, ..., Z_n$ are autocorrelated. Random variables $Y_1, ..., Y_n$ are generated as i.i.d. $N(0, 1)$, and $Z_j = Y_j + Z_{j-1}, j = 2, .., n, Z_1 = Y_1$.

Comparability on Unknown Covariates

Table 2 : *Simulated* $P(|\bar{Z}_A - \bar{Z}_B| > 0.4)$ *for three different types of covariate streams,* $n = 50,\ 100,000$ *replications.*

Procedure	Model (1)	Model (2)	Model (3)
CR	0.162	0.363	0.312
RAR	0.157	0.357	0.308
TBD	0.157	0.541	0.329
RBD ($K = 3$)	0.156	0.169	0.246
RBD ($K = 10$)	0.157	0.205	0.291
Smith ($\rho = 1$)	0.159	0.285	0.131
Smith ($\rho = 5$)	0.157	0.217	0.273
BCD ($p = 2/3$)	0.158	0.236	0.291
Big stick ($b = 3$)	0.157	0.218	0.274
BCDII($p = 2/3, b = 3$)	0.157	0.191	0.277
ABCD($a = 10$)	0.157	0.193	0.248

Comparability on Unknown Covariates

- All procedures work as intended for the i.i.d. sequence. In the case that really matters (time trends and autocorrelation), the random block design and the ABCD work the best.
- We can see the similarity between the balancing on unobserved covariates and balancing on treatment assignments.
- Balancing on baseline covariates is often considered to be so important that a Table 1 is presented in clinical trials papers with multiple hypothesis tests that presume that non-significance implies that the randomization worked. Some people use covariate-adjusted randomization (e.g., minimization, dynamic allocation) to ensure that Table 1 looks good. Senn (1994) has a beautiful and undercited paper debunking the folklore of Table 1.
- What we show here is that, while randomization works asymptotically, for small samples, some procedures may work better than others.

Predictability

- The second great property of randomization is that it provides sequences that are unpredictable. In fact, only complete randomization does this perfectly, with a probability of $1/2$ of assignment to A for all n patients.
- In restricted randomization, the unconditional probability of assignment to A is $1/2$, but the conditional probability, conditional on the previous treatment assignments, may not be. In extreme cases, some elements of a permuted block design may have conditional probability 1.
- One metric of predictability, then, is $\rho_{PRED} = \sum_{j=1}^{n} E|\phi_j - 1/2|$, the sum of the expected differences between the conditional and unconditional assignment probabilities.

Predictability

- Dupin-Spriet, et al., 2004 defined predictability as the proportion of treatment assignments made with probability 1. However, this ignores the higher predictability of assignment with probability 0.9, for instance.
- Some have argued that there is no reason to be concerned about predictability of a sequence with allocation concealment and masking. Berger's (2005) book on selection bias and associated papers, gives multiple examples of selection bias in practice. Senn (2005) describes selection bias in the context of clinical trials that are "randomised by intention but not execution."
- Selection bias can lead to biased treatment effect estimators, inflated type I error, and covariate imbalances.

Predictability

Table 3 : *Simulated* ρ_{PRED}, $n = 50$, *based on* $100,000$ *replications.*

Complete	-0.10
Smith ($\rho = 1$)	3.00
Smith ($\rho = 5$)	6.54
BCD($p = 2/3$)	6.09
Big Stick ($b = 3$)	3.95
BCDII ($p = 2/3, b = 3$)	7.00
ABCD($a = 10$)	6.00
Maximal Procedure ($b = 3$)	6.61

Predictability

Table 4 : *Simulated* ρ_{PRED}, $n = 50$, *based on* $100,000$ *replications.*

Hadamard	6.77
Permuted Block Design ($b = 1$, RAR)	12.50
Permuted Block Design ($b = 2$, RAR)	10.16
Permuted Block Design ($b = 2$, TBD)	8.99
Permuted Block Design ($b = 4$, RAR)	8.04
Permuted Block Design ($b = 4$, TBD)	6.56
Random Block Design ($K = 3$, RAR)	10.01
Random Block Design ($K = 3$, TBD)	8.94
Random Block Design ($K = 10$, RAR)	6.71
Random Block Design ($K = 10$, TBD)	5.16

Predictability

- Predictability and balance are competing objectives. The procedures that are the best at balancing are the most predictable. For an extreme example, a deterministic alternating sequence is perfectly balanced, but perfectly predictable. Complete randomization is the worst procedure with respect to balancing, but has no predictability.
- The maximal procedure was specifically designed to be less predictable, but it has about the same predictability as the permuted block design with $b = 4$. Other restricted randomization procedures, such as Smith's design (Wei's urn design) perform much better, do not require terminal balance, and are easier to generate.
- The random block design is often used to eliminate predictability, but the predictability is still present, and is about the same as the permuted block design with fixed blocks equal to the average block size. Blocks filled with the truncated binomial design are better less predictable.

Trade-off Plots

- Note that $Var(D_n)$ is between $(0, n)$, and ρ_{PRED} is between $(0, n/2)$. We can put them on the same $(0, 1)$ scale by dividing by the upper limit. We can then plot them against each other. If both scales are considered equivalent, and each criterion considered equally important, the procedure falling closest to the origin can be considered the better procedure. Similar "trade-off" plots were suggested by Zhao, et al. (2011), Flournoy, et al. (2013), and Atkinson (2014).
- And the winner is...

Trade-off Plots

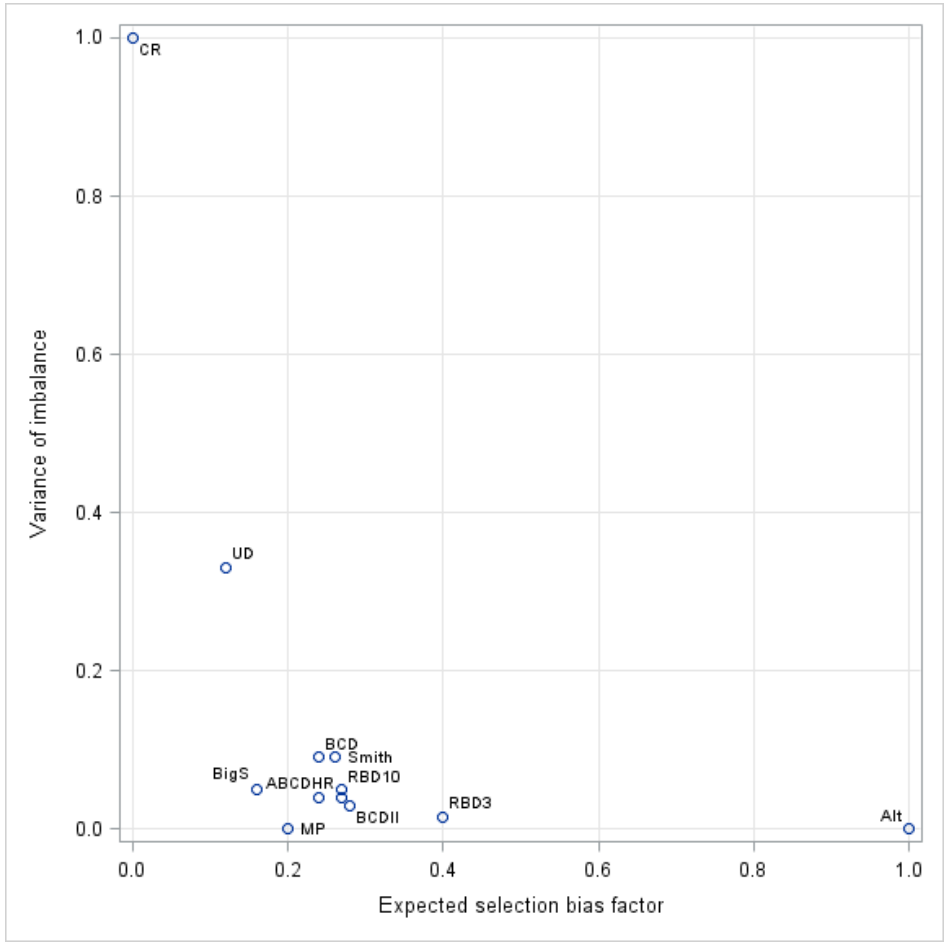


Figure 2 : *Trade-off between balance variability and predictability for 11 procedures,* $n = 50$, $100,000$ *replications*

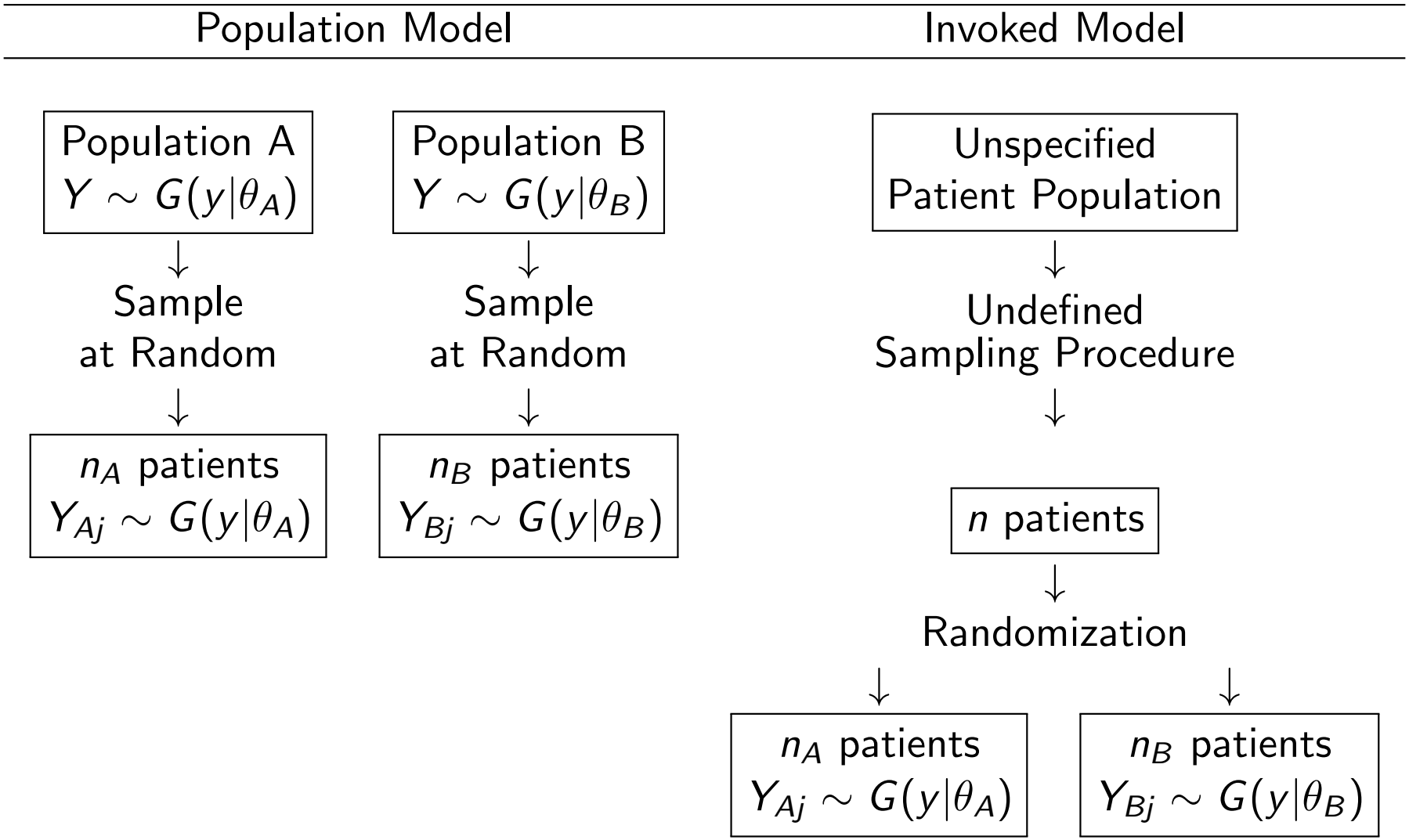
The Big Stick Design

Table of Contents

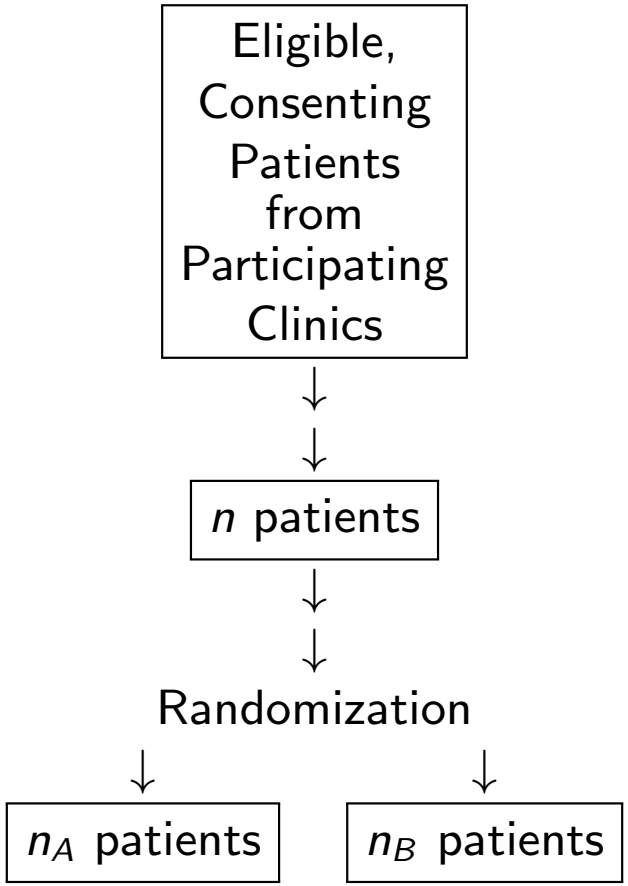
Randomization as a Basis for Inference

- The third 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.
- Randomization tests are particularly useful for small clinical trials, where standard asymptotic tests may not apply.

The Population Model



The Randomization Model



Randomization Tests

- Under the null hypothesis of no treatment effect, the arrangement of patient responses depend only on the way in which they were randomized. If we were to "re-randomize" them many times, the treatment effect should not be affected. The only way in which a treatment effect can exhibit itself is if the results change when we "re-randomize".
- An exact test requires the enumeration of all possible sequences from the chosen randomization procedure, along with the associated probabilities of that treatment.

Randomization Tests

- The p-value is simply the sum of probabilities of sequences that give a treatment effect as extreme or more extreme than the one observed in the trial; i.e., $p = \sum_{l=1}^{\Omega} I(|S_l| \geq |S_{obs.}|) \Pr(L = l)$, where L is the randomization sequence, Ω is the set of all possible randomization sequences, S_l is the treatment effect metric (test statistic) for the lth sequence, and $S_{obs.}$ is the treatment effect (test statistic) that was observed in the trial.
- Only under complete randomization are the sequences equiprobable. For restricted randomization, the probability of each sequence must be calculated. Fortunately we don't have to do that anymore.
- Note that the treatment effect metric can be anything that make sense: difference in means, difference in proportions, linear rank test (Wilcoxon, logrank, logrank with censoring),....

Monte Carlo Randomization Tests

- Instead of exact enumeration, we can randomly generate M sequences of length n, and then compute a Monte Carlo p-value as the proportion of sequences generated that yield a test statistics that is as extreme or more extreme than the observed test statistic; i.e., $\hat{p} = \sum_{l=1}^{M} I(|S_l| \geq |S_{obs.}|)/M$.
- This is a consistent estimator of the randomization p-value, but the consistency is for large M, and has nothing to do with n. Plamadeala and Rosenberger (2012) do some calculations that show that 15,000 sequences is sufficient to estimate even small p-values accurately.
- Plamadeala and Rosenberger (2012) also show how to conduct randomization tests with respect to a conditional reference set, using only those sequences that achieve the same terminal treatment numbers (i.e., conditional on $N_A(n) = n_{A,obs.}$).

Preserving Error Rates

- 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.

Preserving 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
$UD(0,1)$	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

Trade-off Between Predictability and Type II Error

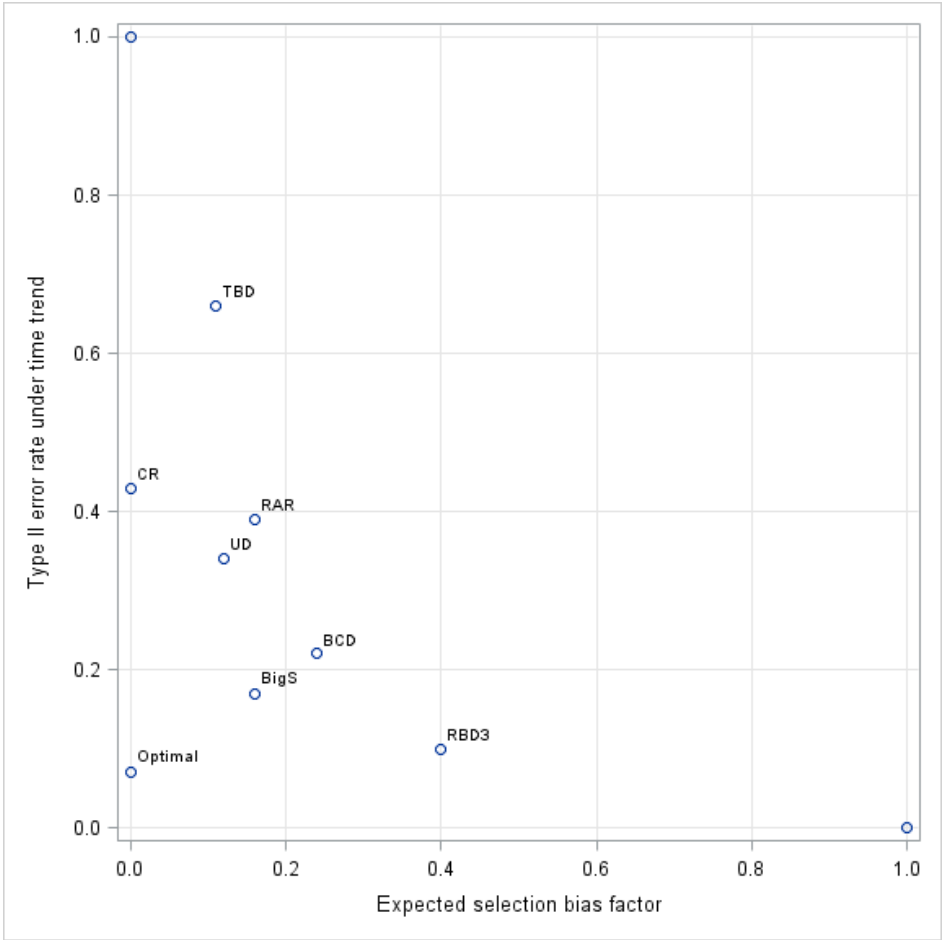


Figure 3 : *Trade-off between predictability and type II error under a linear time trend for 8 procedures,* $n = 50$, $10,000$ *replications*

Randomization Tests for Regression Models

- Gail, et al. (1988) recognized that treatment effects from adjusted regression models can be tested using randomization-based inference. Under the null hypothesis, the treatment effect is 0, and the ranked model residuals should be randomly distributed across all possible randomization sequences. Hence, a randomization test can be performed on the ranked residuals, and a significant randomization p-value is then evidence against the null hypothesis.
- Parwen, et al. (2013) used this approach under generalized linear models, survival models, and GEE models, "re-randomizing" the residuals using Monte Carlo methods.
- Interestingly, while the model may be parametric, the inference procedure is completely nonparametric. They demonstrate that under misspecified models, while the model parameter estimators may be biased, the size of the randomization test is unaffected.

Randomization Tests for Regression Models

- This framework makes available randomization-based inference for virtually every possible type of outcome analysis that may be encountered in randomized clinical trials:
  1. Treatment effects with continuous, ordinal, categorical, time-to-event, and repeated measures outcomes.
  2. Covariate-adjusted treatment effects with continuous, ordinal, categorical, time-to-event, and repeated measures outcomes.
- Such tests can be used as the primary analysis, a complementary analysis, or a sensitivity analysis along with standard parametric tests (that assume random sampling!).
- Many times, randomization tests give the same answer as parametric tests. When samples are small, parametric tests may not be as accurate as we often assume.

Table of Contents

Conclusions

- For 70 years clinical trialists have known that randomization reduces bias due to predictability, induces comparability among important known and unknown covariates, and provides a basis for inference.
- Randomization is often done in a nonchalant, haphazard manner, with no consideration for finding the best procedure for a particular clinical trial, and ignoring the randomization in the analysis.
- In small clinical trials, which may be necessary in rare diseases, typical asymptotic assumptions may not hold.
- We have described different randomization procedures, investigated their properties in small samples, and discussed how to conduct randomization-based inference.

Conclusions

Randomization Matters!!