A roadmap to using randomization in clinical trials

Vance Berger
National Institutes of Health
Louis Bour
Boehringer Ingelheim (Germany)
Kerstine Carter
Boehringer Ingelheim (United States)
Jonathan Chipman
University of Utah
Colin Everett
University of Leeds
Nicole Heussen
Sigmund Freud University Vienna
Catherine Hewitt
University of York
Ralf-Dieter Hilgers
RWTH Aachen University
Y. Abigail Luo
United States Food and Drug Administration
Jone Renteria
University of Barcelona
Yevgen Ryeznik
AstraZeneca (Sweden)
Oleksandr Sverdlov (alex.sverdlov@novartis.com)
Novartis (United States)
Diane Uschner
George Washington University

Research Article

Keywords: balance, randomization-based test, restricted randomization design, validity

Posted Date: January 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-135735/v1

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on August 16th, 2021. See the published version at https://doi.org/10.1186/s12874-021-01303-z.

A roadmap to using randomization in clinical trials

Randomization Innovative Design Scientific Working Group

Vance W. Berger [1]
Louis Joseph Bour [2]
Kerstine Carter [3]
Jonathan J. Chipman [4]
Colin Everett [5]
Nicole Heussen [6, 7]
Catherine Hewitt [8]
Ralf-Dieter Hilgers [6]
Y. Abigail Luo [9]
Jone Renteria [10, 11]
Yevgen Ryeznik [12]
Oleksandr Sverdlov [13,*]
Diane Uschner [14]

[1] National Institutes of Health, Bethesda MD, USA
[2] Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany
[3] Boehringer-Ingelheim Pharmaceuticals Inc., Ridgefield CT, USA
[4] University of Utah School of Medicine, Salt Lake City UT, USA
[5] University of Leeds, Leeds, United Kingdom
[6] RWTH Aachen University, Aachen, Germany
[7] Medical School, Sigmund Freud University, Vienna, Austria
[8] University of York, York, United Kingdom
[9] Food and Drug Administration, Rockville MD, USA
[10] Open University of Catalonia (UOC) and the University of Barcelona (UB), Spain
[11] Department of Human Development and Quantitative Methodology, University of Maryland, College Park MD, USA
[12] Early Biometrics & Statistical Innovations, Data Science & Artificial Intelligence, R&D, AstraZeneca, Gothenburg, Sweden
[13] Novartis Pharmaceuticals Corporation, East Hanover NJ, USA
[14] George Washington University, Washington DC, USA

***Correspondence:**
alex.sverdlov@novartis.com

Abstract

Background: Randomization is the foundation of any clinical trial involving treatment comparison. It helps mitigate selection bias, promotes similarity of treatment groups with respect to important known and unknown confounders, and contributes to the validity of statistical tests. Various restricted randomization procedures with different probabilistic structures and different statistical properties are available.

Methods: We survey available restricted randomization procedures for sequential allocation of subjects in a randomized, comparative, parallel group clinical trial with equal (1:1) allocation. We explore statistical properties of these procedures, including balance/randomness tradeoff, type I error rate and power. We perform head-to-head comparisons of different procedures through simulation under various experimental scenarios, including cases when common model assumptions are violated. We also provide some real-life clinical trial examples to illustrate the thinking process for selecting a randomization procedure for implementation in practice.

Results: Restricted randomization procedures targeting 1:1 allocation vary in the degree of balance/randomness they induce, and more importantly, they vary in terms of validity and efficiency of statistical inference when common model assumptions are violated (e.g. when outcomes are affected by a linear time trend; measurement error distribution is misspecified; or selection bias is introduced in the experiment). Some procedures are more robust than others. Covariate-adjusted analysis may be essential to ensure validity of the results. Special considerations are required when selecting a randomization procedure for a clinical trial with very small sample size.

Conclusions: The choice of randomization design, data analytic technique (parametric or nonparametric), and analysis strategy (randomization-based or population model-based) are all very important considerations. Randomization-based tests are robust and valid alternatives to likelihood-based tests and should be considered more frequently by clinical investigators.

Keywords: balance, randomization-based test, restricted randomization design, validity

Background

Various research designs can be used to acquire scientific medical evidence. The randomized controlled trial (RCT) has been recognized as the most credible research design for investigations of the clinical effectiveness of new medical interventions [1, 2]. Evidence from RCTs is widely used as a basis for submissions of regulatory dossiers in request of marketing authorization for new drugs, biologics, and medical devices. Three important methodological pillars of the modern RCT include blinding (masking), randomization, and the use of control group [3].

While RCTs provide the highest standard of clinical evidence, they are laborious and costly, in terms of both time and material resources. There are alternative designs, such as observational studies with either a cohort or case-control design, and studies using real world evidence (RWE). In the era of big data, the sources of clinically relevant data are increasingly rich and include electronic health records, data collected from wearable devices, etc. RWE studies can be implemented rapidly and relatively easily. But how credible are the results from such studies?

In 1980, D. P. Byar issued warnings and highlighted potential methodological problems with comparison of treatment effects using observational databases [4]. Many of these issues still persist and actually become paramount during the ongoing COVID-19 pandemic when global scientific efforts are made to find safe and efficacious vaccines and treatments as soon as possible. Recently, two top medical journals, the New England Journal of Medicine and the Lancet, retracted two COVID-19 studies that relied on observational registry data [5, 6]. The retractions were made at the request of the authors who were unable to ensure reproducibility of the results [7]. Undoubtedly, such cases are harmful in many ways. The already approved drugs may be wrongly labeled as “toxic” or “inefficacious”, and the reputation of the drug developers could be blemished or destroyed. Therefore, the highest standards for clinical evidence are now needed more than ever. When treatment effects are modest, yet still clinically meaningful, a double-blind, randomized, controlled clinical trial design helps detect these differences while adjusting for possible confounders and adequately controlling the chances of both false positive and false negative findings.

Randomization in clinical trials has been an important area of research in biostatistics since the pioneering work of A. Bradford Hill in the 1940’s and the first published randomized trial comparing streptomycin with a non-treatment control [8]. Statisticians around the world have

worked intensively to elaborate the value, properties, and refinement of randomization procedures with an incredible record of publication [9]. In particular, a recent EU-funded project (www.IDeAl.rwth-aachen.de) on innovative design and analysis of small population trials has "randomization" as one work package. In 2020, a group of trial statisticians around the world from different sectors formed a subgroup of the Drug Information Association (DIA) Innovative Designs Scientific Working Group (IDSWG) to raise awareness of the full potential of randomization to improve trial quality, validity and rigor (https://randomization-working-group.rwth-aachen.de/).

The aims of the current paper are three-fold. First, we describe major recent methodological advances in randomization, including different restricted randomization designs that have superior statistical properties compared to some widely used procedures such as permuted block designs. Second, we discuss different types of experimental biases in clinical trials and explain how a carefully chosen randomization design can mitigate risks of these biases. Third, we provide a systematic template for evaluating different restricted randomization procedures and selecting an "optimal" one for a particular trial. We also showcase application of these ideas through several real life RCT examples.

The target audience for this paper are clinical investigators and biostatisticians who are tasked with the design, conduct, analysis, and interpretation of clinical trial results, as well as regulatory and scientific/medical journal reviewers. Recognizing the breadth of the concept of randomization, in this paper we focus on a randomized, comparative, parallel group clinical trial design with equal (1:1) allocation, which is typically implemented using some restricted randomization procedure, possibly stratified by some important baseline prognostic factor(s) and/or study center. Some of our findings and recommendations are generalizable to more complex clinical trial settings. We shall highlight these generalizations and outline additional important considerations that fall outside the scope of the current paper.

Methods

What is randomization and what are its virtues in clinical trials?

Randomization is an essential component of an experimental design in general and clinical trials in particular. Its history goes back to R. A. Fisher and his classic book "The Design of Experiments" [10]. Implementation of randomization in clinical trials is due to A. Bradford Hill

who designed the first randomized clinical trial evaluating the use of streptomycin in treating tuberculosis in 1946 [8, 11, 12].

References [13, 14] provide good summaries of the rationale and justification for the use of randomization in clinical trials. The randomized controlled trial (RCT) has been referred to as "the worst possible design (except for all the rest)" [15], indicating that the benefits of randomization should be evaluated in comparison to what we are left with if we do not randomize. Observational studies suffer from a wide variety of biases that are built into the design, and do not vanish when we pretend that we can address them adequately with modeling assumptions. The intent of randomization is to create comparable treatment groups, and while this objective is not always met, it can be almost guaranteed to fail with observational studies.

The RCT in the medical field has several features that distinguishes it from experimental designs in other fields, such as agricultural experiments. In the RCT, the experimental units are humans, and in the medical field often diagnosed with a potentially fatal disease. These subjects are sequentially enrolled for participation in the study at selected study centers, which have relevant expertise for conducting clinical research. Many contemporary clinical trials are run globally, at multiple research institutions. The recruitment period may span several months or even years, depending on a therapeutic indication and the target patient population. Patients who meet study eligibility criteria must sign the informed consent, after which they are enrolled into the study and, for example, randomized to either experimental treatment E or the control treatment C according to the randomization sequence. In this setup, the choice of the randomization design must be made judiciously, to protect the study from experimental biases and ensure validity of clinical trial results.

The first virtue of randomization is that, in combination with allocation concealment and masking, it helps mitigate selection bias due to an investigator's potential to selectively enroll patients into the study [16]. A non-randomized, systematic design such as a sequence of alternating treatment assignments has a major fallacy: an investigator, knowing an upcoming treatment assignment in a sequence, may enroll a patient who, in their opinion, would be best suited for this treatment. Consequently, one of the groups may contain a greater number of "sicker" patients and the estimated treatment effect may be biased. Systematic covariate imbalances may increase the probability of false positive findings and undermine the integrity of the trial. While randomization

alleviates the fallacy of a systematic design, it does not fully eliminate the possibility of selection bias (unless we consider complete randomization for which each treatment assignment is determined by a flip of a coin, which is rarely, if ever used in practice [17]). Commonly, RCTs employ restricted randomization procedures which sequentially balance treatment assignments while maintaining allocation randomness. To minimize potential for selection bias, one should avoid overly restrictive randomization schemes such as permuted block design with small block sizes, as this is very similar to alternating treatment sequence.

The second virtue of randomization is its tendency to promote similarity of treatment groups with respect to important known, but even more importantly, unknown confounders. If treatment assignments are made at random, then by the law of large numbers, the average values of patient characteristics should be approximately equal in the experimental and the control groups, and any observed treatment difference should be attributed to the treatment effects, not the effects of the study participants [18]. However, one can never rule out the possibility that the observed treatment difference is due to chance, e.g. as a result of random imbalance in some patient characteristics [19]. Despite that random covariate imbalances can occur in clinical trials of any size, such imbalances do not compromise the validity of statistical inference. An unbiased treatment comparison can be achieved through proper covariate-adjusted analysis [20]. It should be noted that some randomization designs, such as covariate-adaptive randomization procedures, can achieve very tight balance of covariate distributions between treatment groups [21]. While we address randomization within pre-specified stratifications, we do not address more complex covariate- and response-adaptive randomization in this paper.

Finally, randomization plays an important role in statistical analysis of the clinical trial. The most common approach to inference following the RCT is the *invoked population model* [9]. With this approach, one posits that there is an unspecified target population of patients with the disease, from which n eligible subjects are sampled for the study and are randomized to the treatment groups. Within each group, the responses are assumed to be independent and identically distributed (i.i.d.), and inference on the treatment effect is performed using some standard statistical methodology, e.g. a two sample t-test for normal outcome data. The added value of randomization is that it makes the assumption of i.i.d. errors more feasible compared to a non-randomized study because it introduces a real element of chance in the allocation of patients.

An alternative approach is the *randomization model*, in which the experimental randomization itself forms the basis for statistical inference [9]. Under the null hypothesis of the equality of treatment effects, individual outcomes are not affected by treatment and are regarded as fixed. Treatment assignments are permuted in all possible ways consistent with the randomization procedure actually used in the trial. The randomization-based p-value is the sum of randomization sequences that yield the treatment difference the same or more extreme than the one actually observed in the trial. A randomization-based test can be a useful supportive analysis, free of assumptions of parametric tests and protective against spurious significant results that may be caused by temporal trends [14, 22].

It is important to note that Bayesian inference has also become a common statistical analysis in RCTs [23]. Although the inferential framework relies upon subjective probabilities, a study analyzed through a Bayesian framework still relies upon randomization for the other aforementioned virtues [24]. Hence, the randomization considerations discussed herein have broad application.

What types of randomization methodologies are available?

Randomization is not a single methodology, but a very broad class of design techniques for the RCT [9]. In this paper, we consider only randomization designs for sequential enrollment clinical trials with equal (1:1) allocation in which randomization is not adapted for covariates and/or responses. The simplest procedure for an RCT is complete randomization (CRD) for which each subject's treatment is determined by a flip of a fair coin [25]. CRD provides no potential for selection bias but it can result, with non-negligible probability, in deviations from the 1:1 allocation ratio, especially in small samples. In practice, some restrictions on randomization are made to achieve balanced allocation. Such randomization designs are referred to as *restricted randomization* procedures [26, 27].

Suppose n (even) subjects are to be randomized sequentially between treatments E and C. Two basic designs that equalize the final treatment numbers are the random allocation rule (Rand) and the truncated binomial design (TBD) [28]. For Rand, any sequence of exactly $n/2$ E's and $n/2$ C's is equally likely. For TBD, treatment assignments are made with probability 0.5 until one of the treatments receives its quota of $n/2$ subjects; thereafter all remaining assignments are made deterministically to the opposite treatment.

A disadvantage of both Rand and TBD is that they aim at the final balance, whereas at intermediate steps it is still possible to have substantial imbalances, especially if n is large. A long run of a single treatment in a sequence may be problematic if there is a time drift in some important covariate, which can lead to chronological bias [29]. To mitigate this risk, one can further restrict randomization so that treatment assignments are balanced over time. The most common approach is the permuted block design (PBD) [30], for which treatment assignments are made using either Rand or TBD in blocks of size $2b$ (b is some small positive integer), with exactly b allocations to each of the treatments E and C. One challenge with PBD is the choice of the block size. If $b = 1$, then every pair of allocations is balanced, but every even allocation is deterministic. Larger block sizes increase allocation randomness. The use of variable block sizes has been suggested [31]; however, PBDs with variable block sizes are also quite predictable [32]. Another problematic feature of the PBD is that it forces periodic return to perfect balance, which may be unnecessary from the statistical efficiency perspective and may increase the risk of prediction of upcoming allocations.

Better alternatives to PBD, called the maximum tolerated imbalance (MTI) procedures have been developed [33]. These procedures provide stronger encryption of the randomization sequence while controlling treatment imbalance at a pre-defined level throughout the experiment. A general MTI procedure specifies a certain boundary for treatment imbalance, say $b > 0$, that cannot be exceeded. If, at a given allocation step the absolute value of imbalance is equal to b, then one next allocation is deterministically forced toward balance. This is in contrast to PBD which, after reaching the target quota of allocations for either treatment within a block, forces all subsequent allocations to achieve perfect balance at the end of the block. Several MTI procedures with different probabilistic structures have been proposed in the literature [34–41]. These designs ensure control of treatment imbalance within pre-specified limits and are more immune to selection bias than PBD [42, 43].

An important class of restricted randomization procedures is biased coin designs (BCDs). Starting with the seminal work of Efron [44], BCDs have been a hot research topic in biostatistics for almost 50 years. Efron's BCD is very simple: at any allocation step, if treatment numbers are balanced, the next assignment is made with probability 0.5; otherwise, the underrepresented treatment is assigned with probability p, where $p \in (0.5,1]$ is a fixed and pre-specified parameter

that determines the tradeoff between balance and randomness. Note that $p = 1$ corresponds to PBD with block size 2. If we set $p < 1$ (e.g. $p = 2/3$), then the procedure has no deterministic assignments and treatment allocation will be concentrated around 1:1 with high probability [44]. Several extensions of Efron's BCD providing better tradeoff between treatment balance and allocation randomness have been proposed [45–49] and a comprehensive comparison of different BCDs has been published [50].

Finally, urn models provide a useful mechanism for RCT designs [51]. Some notable randomized urn designs for balancing treatment assignments are available [39, 40, 52– 55]. Urn designs involve some parameters that can be fine-tuned to obtain randomization procedures with desirable balance/randomness tradeoff [56].

What are the attributes of a good randomization procedure?

A "good" randomization procedure is one that helps successfully achieve the study objective(s). Kalish and Begg [57] state that the major objective of a comparative clinical trial is to provide a precise and valid comparison. To achieve this, the trial design should be such that it: 1) prevents bias; 2) ensures an efficient treatment comparison; and 3) is simple to implement to minimize operational errors. Table 1 elaborates on these considerations, focusing on restricted randomization procedures for 1:1 randomized trials. Before delving into a detailed discussion, let us introduce some important definitions.

Table 1: Considerations for the choice of a restricted randomization procedure

Objective	Desired feature(s) of a randomization procedure
Mitigate potential for selection bias	• A procedure should have high degree of randomness (low proportion of deterministic assignments and low correct guess probability).
Mitigate potential for chronological bias	• A procedure should balance treatment assignments over time.
Valid and efficient treatment comparison	• A procedure should have established statistical properties, provide strong control of false positive rate and yield unbiased, low variance estimates of the treatment difference. • A procedure should attain the chosen target allocation ratio with high probability.
Ease of implementation	• Validated statistical software for implementing a randomization procedure must be in place.

Following [9], a *randomization sequence* is a random vector $\boldsymbol{\delta}_n = (\delta_1, \ldots, \delta_n)$, where $\delta_i = 1$, if the ith subject is assigned to treatment E or $\delta_i = 0$, if the ith subject is assigned to treatment C. A *restricted randomization procedure* can be defined by specifying a probabilistic rule for the treatment assignment of the $(i+1)$st subject, δ_{i+1}, given the past allocations $\boldsymbol{\delta}_i$ for $i \geq 1$. Let $N_E(i) = \sum_{j=1}^{i} \delta_j$ and $N_C(i) = i - N_E(i)$ denote the numbers of subjects assigned to treatments E and C, respectively, after i allocation steps. Then $D(i) = N_E(i) - N_C(i)$ is *treatment imbalance* after i allocations. For any $i \geq 1$, $D(i)$ is a random variable whose probability distribution is determined by the chosen randomization procedure.

Balance and randomness

Treatment balance and allocation randomness are two competing requirements in the design of an RCT. Restricted randomization procedures that provide a good tradeoff between these two criteria are desirable in practice.

Consider a trial with sample size n. The absolute value of imbalance, $|D(i)|$ $(i = 1, \ldots, n)$, provides a measure of deviation from equal allocation after i allocation steps. $|D(i)| = 0$ indicates that the trial is perfectly balanced. One can also consider $\Pr(|D(i)| = 0)$, the probability of achieving exact balance after i allocation steps. In particular $\Pr(|D(n)| = 0)$ is the probability that the final treatment numbers are balanced. Two other useful summary measures are the expected imbalance at the ith step, $E|D(i)|$ and the expected value of the maximum imbalance of the entire randomization sequence, $E\left(\max_{1 \leq i \leq n} |D(i)|\right)$.

Greater forcing of balance implies lack of randomness. A procedure that lacks randomness may be susceptible to selection bias [16]. A classic approach to quantify the degree of susceptibility of a procedure to selection bias is the Blackwell-Hodges model [28]. Let $G_i = 1$ (or 0), if at the ith allocation step an investigator makes a correct (or incorrect) guess on treatment assignment δ_i, given past allocations $\boldsymbol{\delta}_{i-1}$. Then the predictability of the design at the ith step is the expected value of G_i, i.e. $E(G_i) = \Pr(G_i = 1)$. Blackwell and Hodges [28] considered the *expected bias factor*, the difference between expected total number of correct guesses of a given sequence of random assignments and the similar quantity obtained from CR for which treatment assignments are made independently with equal probability: $E(F) = E(\sum_{i=1}^{n} G_i) - n/2$. This quantity is zero for CRD, and it is positive for restricted randomization procedures (greater values indicate higher

expected bias). Matts and Lachin [30] suggested taking *expected proportion of deterministic assignments* in a sequence as another measure of lack of randomness.

In the literature, various restricted randomization procedures have been compared in terms of balance and randomness [50, 58, 59]. For instance, a comprehensive simulation study was performed of 14 restricted randomization procedures with different choices of design parameters, for sample sizes in the range of 10 to 300 [58]. The key criteria were the maximum absolute imbalance and the correct guess probability. The authors found that the performance of the designs was within a closed region with the boundaries shaped by Efron's BCD [44] and the big stick design [37]. The latter procedure with a suitably chosen MTI boundary can achieve very good balance/randomness tradeoff [60].

Statistical properties – validity and efficiency

Validity of a statistical procedure essentially means that the procedure provides correct statistical inference following an RCT. In particular, a chosen statistical test is valid, if it controls the chance of a false positive finding, that is, the pre-specified probability of a type I error of the test is achieved but not exceeded. The strong control of type I error rate is a major prerequisite for any confirmatory RCT. *Efficiency* means high statistical power for detecting meaningful treatment differences (when they exist), and high accuracy of estimation of treatment effects.

Both validity and efficiency are major requirements of any RCT, and both of these aspects are intertwined with treatment balance and allocation randomness. Restricted randomization designs, when properly implemented, provide solid ground for valid and efficient statistical inference. However, a careful consideration of different options can help an investigator to optimize the choice of a randomization procedure for their clinical trial.

Let us start with statistical efficiency. Equal (1:1) allocation frequently maximizes power and estimation precision. To illustrate this, suppose the primary outcomes in the two groups are normally distributed with respective means μ_E and μ_C and standard deviation $\sigma > 0$. Then the variance of an efficient estimator of the treatment difference $\mu_E - \mu_C$ is equal to $V = \frac{4\sigma^2}{n - L_n}$, where $L_n = \frac{|D(n)|^2}{n}$ is referred to as *loss* [61]. Clearly, V is minimized when $L_n = 0$, or equivalently, $D(n) = 0$, i.e. the balanced trial.

When the primary outcome follows a more complex statistical model, optimal allocation may be unequal across the treatment groups; however, 1:1 allocation is still nearly optimal for binary outcomes [62, 63], survival outcomes [64], and possibly more complex data types [65, 66]. Therefore, a randomization design that balances treatment numbers frequently promotes efficiency of the treatment comparison.

As regards inferential validity, it is important to distinguish two approaches to statistical inference after the RCT – an *invoked population* model and a *randomization* model [9]. For a given randomization procedure, these two approaches generally produce similar results when the assumption of normal random sampling (and some other assumptions) are satisfied, but the randomization model may be more robust when model assumptions are violated; e.g. when outcomes are affected by a linear time trend [67, 68]. Another important issue that may interfere with validity is selection bias. Some authors showed theoretically that PBDs with small block sizes may result in serious inflation of the type I error rate under a selection bias model [69–71]. To mitigate risk of selection bias, one should ideally take preventative measures, such as blinding/masking, allocation concealment, and avoidance of highly restrictive randomization designs. However, for already completed studies with evidence of selection bias [73], special statistical adjustments are warranted to ensure validity of the results [74–76].

Implementation aspects

With the current state of information technology, implementation of randomization in the RCT should be straightforward. Validated systems are emerging, and they can handle randomization designs of increasing complexity for clinical trials that are run globally. However, some important points merit consideration.

The first point has to do with how a randomization sequence is generated and implemented. There is a distinction between *advance* and *adaptive* randomization [16]. The former approach includes the following steps: 1) for the chosen randomization design and sample size n, specify the probability distribution on the reference set by enumerating all feasible randomization sequences of length n and their corresponding probabilities; 2) select a sequence at random from the reference set according to the probability distribution; and 3) implement this sequence in the trial. A “weak link” of the described approach is that the sequence of treatment assignments is obtained upfront, and then proper security measures (e.g. blinding/masking) must be in place to

protect confidentiality and maintain study integrity. With the adaptive randomization, a sequence of treatment assignments is generated dynamically as the trial progresses. For many restricted randomization procedures, the randomization rule can be expressed as $\Pr(\delta_{i+1} = 1) = F\{D(i)\}$, where $F\{\cdot\}$ is some non-increasing function of $D(i)$ for any $i \geq 1$. This is referred to as the *Markov property* [77], which makes a procedure easy to implement sequentially. Some restricted randomization procedures, e.g. the maximal procedure [35], do not have the Markov property.

The second point has to do with how the final data analysis is performed. With an invoked population model, the analysis is conditional on the design and the randomization is ignored in the analysis. With a randomization model, the randomization itself forms the basis for statistical inference. A contemporaneous overview of randomization-based inference in clinical trials has been published [14]. Several other papers, provide important technical details on randomization-based tests, including justification for control of type I error rate with these tests [22, 78, 79]. In practice, Monte Carlo simulation can be used to estimate randomization-based p-values [9].

Results

In general, the choice of a randomization procedure is only one part of a clinical trial design. The design of any RCT should start with formulation of the trial objectives and research questions of interest [3, 31]. Suppose it has been decided that the study will be designed as a randomized, two-arm comparative trial with 1:1 allocation, with a fixed sample size n that is pre-determined based on budgetary and statistical considerations to obtain a definitive assessment of the treatment effect via the pre-defined hypothesis testing. Given various restricted randomization procedures that can be applied in sequential enrollment RCTs, it is quite a challenge to find one that is "optimal" for a given trial. In what follows, we provide several examples to illustrate the thinking process for selecting a proper randomization procedure in an RCT.

Which restricted randomization procedures are robust and efficient?

Our first example is a hypothetical RCT in which the primary outcome is assumed to be normally distributed with mean μ_E for treatment E, mean μ_C for treatment C, and common variance σ^2. A total of n subjects are to be randomized equally between E and C, and a two-sample t-test is planned for data analysis. Let $\Delta = \mu_E - \mu_C$ denote the true mean treatment difference. We are interested in testing a hypothesis $H_0: \Delta = 0$ (treatment effects are the same) vs. $H_1: \Delta \neq 0$.

The total sample size n to achieve given power at some clinically meaningful treatment difference Δ_c while maintaining the chance of a false positive result at level α can be obtained using standard statistical methods [80]. For instance, if $\Delta_c/\sigma = 0.95$, then a design with $n = 50$ subjects (25 per arm) provides approximately 91% power of a two-sample t-test to detect a statistically significant treatment difference using 2-sided $\alpha = 5\%$. We shall consider 11 randomization procedures to sequentially randomize $n = 50$ subjects in a 1:1 ratio.

1. Random allocation rule – Rand.
2. Truncated binomial design – TBD.
3. Permuted block design with block size of 2 – PBD(2).
4. Big stick design [37] with MTI=3 – BSD(3).
5. Biased coin design with imbalance tolerance [38] with p=2/3 and MTI=3 – BCDWIT(2/3, 3).
6. Efron's biased coin design [44] with p=2/3 – BCD(2/3).
7. Adjustable biased coin design [49] with a=2 – ABCD(2).
8. Generalized biased coin design (GBCD) with $\gamma = 1$ [45] – GBCD(1).
9. GBCD with $\gamma = 2$ [46] – GBCD(2).
10. GBCD with $\gamma = 5$ [47] – GBCD(5).
11. Completely randomized design – CRD.

These 11 procedures can be grouped into five major types. I) Procedures 1, 2, and 3 achieve exact final balance for a chosen sample size. II) Procedures 4 and 5 ensure that at any allocation step the absolute value of imbalance is capped at MTI=3. III) Procedures 6 and 7 are biased coin designs that sequentially adjust randomization according to imbalance measured as the difference in treatment numbers. IV) Procedures 8, 9, and 10 (CBCD's with γ =1, 2, and 5) are adaptive biased coin designs, for which randomization probability is modified according to imbalance measured as the difference in treatment allocation proportions (larger γ implies greater forcing of balance). V) Procedure 11 (CRD) is the most random procedure that achieves balance for large samples.

Balance/randomness tradeoff

We first compare the procedures with respect to balance and randomness. To quantify treatment imbalance after m allocations, we consider two measures: expected value of absolute imbalance $E|D(m)|$, and expected value of loss, $E(L_m) = \frac{E|D(m)|^2}{m}$ [50, 61]. Importantly, for procedures 1, 2,

and 3, the final imbalance is always zero, thus $E|D(n)| \equiv 0$ and $E(L_n) \equiv 0$), but at intermediate steps we may have $E|D(m)| > 0$ and $E(L_m) > 0$, for $1 \leq m < n$. For procedures 4 and 5 with MTI=3, $E(L_m) \leq 9/m$. For procedures 6 and 7, as n increases $E(L_n)$ tends to zero [49]. For procedures 8, 9, 10, and 11, as n increases, $E(L_n)$ tends to the positive constants 1/3, 1/5, 1/11, and 1, respectively [47]. We take the cumulative average loss after n allocations as an aggregate measure of imbalance: $Imb(n) = \frac{1}{n}\sum_{i=1}^{n} E(L_i)$, which takes values in the 0–1 range.

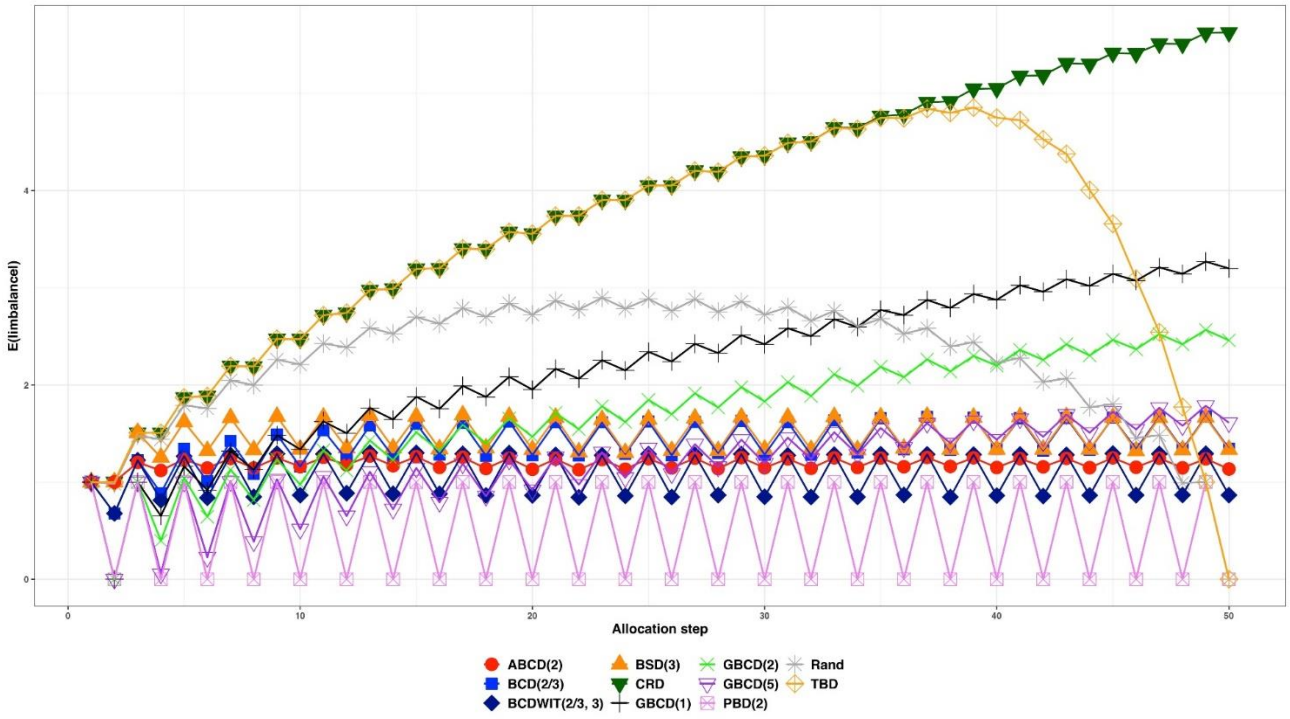
To measure lack of randomness, we consider the *forcing index* [47, 81], $FI(n) = \frac{\sum_{i=1}^{n} E|\phi_i - 0.5|}{n/4}$, where $E|\phi_i - 0.5|$ is the expected deviation of the conditional probability of treatment E assignment at the ith allocation step (ϕ_i) from the unconditional target value (0.5). Note that $FI(n)$ also takes values in the 0–1 range. At the one extreme, we have CRD for which $FI(n) \equiv 0$ (because for CRD $\phi_i = 0.5$ for any $i \geq 1$). At the other extreme, we have PBD(2) for which every odd allocation is made with probability 0.5, and every even allocation is deterministic (made with probability 0 or 1). For PBD(2), there are exactly $n/2$ pairs of allocations, and so $\sum_{i=1}^{n} E|\phi_i - 0.5| = 0.5 \cdot n/2 = n/4$, which implies that $FI(n) = 1$ for PBD(2). For all other restricted randomization procedures one has $0 < FI(n) < 1$.

A "good" randomization procedure should have low values of both loss and forcing index. Different randomization procedures can be compared graphically by plotting $Imb(n)$ vs. $FI(n)$ for the chosen n (in our example $n = 50$). As a balance/randomness tradeoff metric, one can calculate the quadratic distance to the origin (0,0): $d(n) = \sqrt{\{Imb(n)\}^2 + \{FI(n)\}^2}$, and the randomization designs can then be ranked such that designs with lower values of $d(n)$ are preferable.

We ran a simulation study of 11 randomization procedures for an RCT with $n = 50$ subjects. Monte Carlo average values of absolute imbalance, loss, $Imb(m)$, $FI(m)$, and $d(m)$ were calculated for each intermediate allocation step ($m = 1, \ldots, 50$), based on 10,000 simulation runs.

Figure 1 is a plot of expected absolute imbalance vs. allocation step. CRD, GBCD(1), and GBCD(2) show increasing imbalance patterns. For TBD and Rand, the final imbalance (when $n = 50$) is zero; however, at intermediate steps is can be quite large. For other designs, absolute imbalance is expected to be below 2 at any allocation step up to $n = 50$. Note a periodic pattern of PBD(2): imbalance is 0 (or 1) for any even (or odd) allocation.

Figure 1: Simulated expected absolute imbalance vs. allocation step for 11 restricted randomization procedures for n=50.



Note: PBD(2) has a forced periodicity absolute imbalance of 0, which distinguishes it from an MTI procedure

Table 2 shows the ranking of the 11 designs with respect to the overall performance metric $d(n) = \sqrt{\{Imb(n)\}^2 + \{FI(n)\}^2}$ for $n = 50$. BSD(3), GBCD(2) and GBCD(1) are the top three procedures, whereas PBD(2) and CRD are at the bottom of the list.

Table 2: Ranking of 11 restricted randomization procedures with respect to balance / randomness tradeoff for a trial with n=50 subjects.

Rank	Design	FI(n)	Imb(n)	d(n)
1	BSD(3)	0.316	0.226	0.389
2	GBCD(2)	0.344	0.220	0.409
3	GBCD(1)	0.240	0.341	0.417
4	ABCD(2)	0.419	0.170	0.452
5	GBCD(5)	0.522	0.121	0.536
6	BCD(2/3)	0.487	0.233	0.540
7	BCDWIT(2/3, 3)	0.560	0.149	0.579
8	Rand	0.318	0.505	0.597
9	TBD	0.225	0.868	0.896
10	PBD(2)	1	0.052	1.001
11	CRD	0	1.014	1.014

Figure 2 is a plot of $FI(n)$ vs. $Imb(n)$ for $n = 50$. One can see the two extremes: CRD that takes the value (0,1), and PBD(2) with the value (1,0). The other nine designs are closer to (0,0).

Figure 2: Simulated forcing index (x-axis) vs. aggregate expected loss (y-axis) for 11 restricted randomization procedures for n=50.

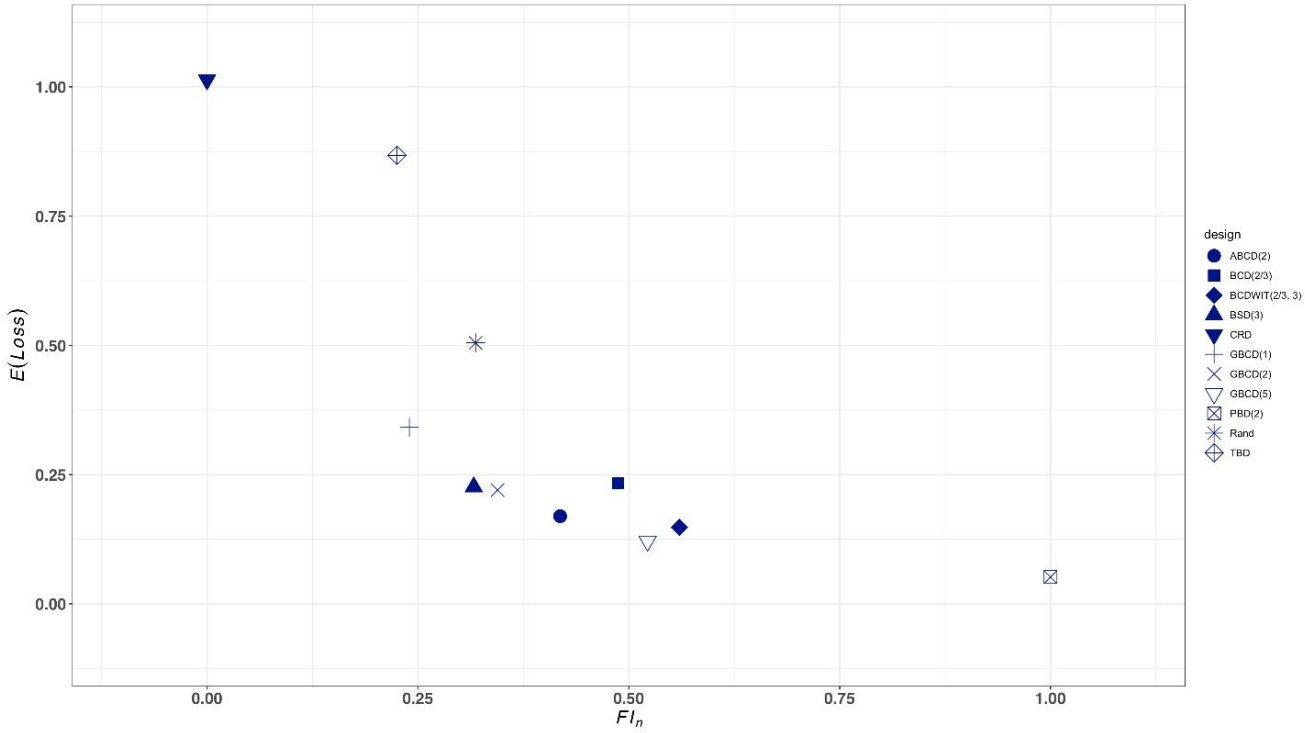
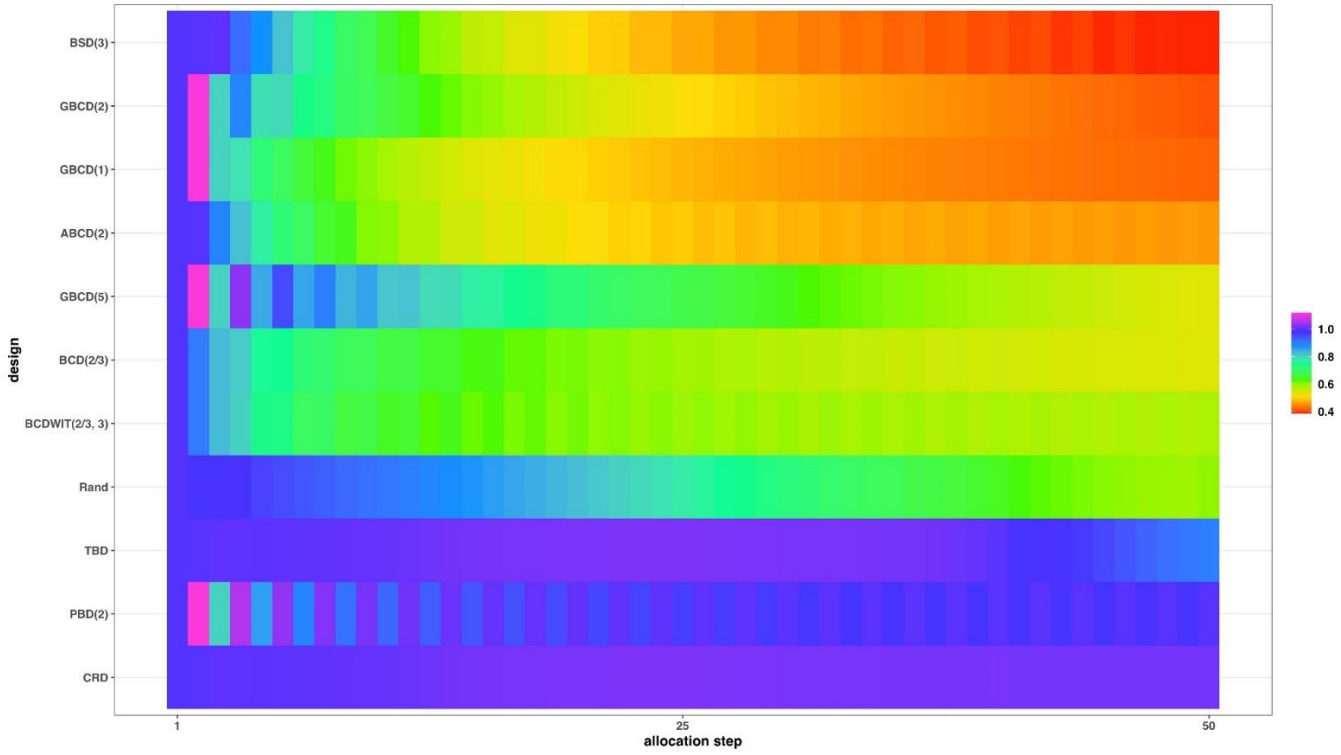


Figure 3 is a heat map plot of the metric $d(m)$ for $m = 1, \ldots, 50$. BSD(3) seems to provide overall best tradeoff between randomness and balance throughout the study.

Figure 3: Heat map plot of the balance/randomness tradeoff metric $\boldsymbol{d(m) = \sqrt{\{Imb(m)\}^2 + \{FI(m)\}^2}}$ vs. allocation step ($\boldsymbol{m = 1, \ldots, 50}$) for 11 restricted randomization procedures.



Inferential characteristics: type I error rate and power

A more rigorous approach to compare the merits of different randomization procedures is to study their inferential characteristics such as type I error rate and power under different experimental conditions. This requires generation of individual outcome data from some plausible models, including scenarios when some common assumptions are violated. We assume that for the ith subject, the data generating mechanism for the outcome Y_i (conditional on the treatment assignment δ_i) is

$$Y_i = \delta_i \mu_E + (1 - \delta_i)\mu_C + u_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where u_i is an unknown term associated with the ith patient and ε_i's are i.i.d. measurement errors. We shall explore the following four models:

- ***M1: Normal random sampling*:** $u_i \equiv 0$ and $\varepsilon_i \sim$ i.i.d. N(0,1), $i = 1, \dots, n$. This corresponds to a standard setup for a two-sample t-test under a population model.
- ***M2: Linear trend*:** $u_i = \frac{5i}{n+1}$ and $\varepsilon_i \sim$ i.i.d. N(0,1), $i = 1, \dots, n$. In this model, the outcomes are affected by a linear trend over time [67].
- ***M3: Cauchy errors*:** $u_i \equiv 0$ and $\varepsilon_i \sim$ i.i.d. Cauchy(0,1), $i = 1, \dots, n$. In this setup, we have a misspecification of the distribution of measurement errors.
- ***M4: Selection bias*:** $u_{i+1} = \nu \cdot 1\{D(i) < 0\} - \nu \cdot 1\{D(i) > 0\}$, $i = 0, \dots, n-1$, with the convention $D(0) = 0$. Here, $\nu > 0$ is the "bias effect" (we set $\nu = 0.5$) and $1\{\cdot\}$ is the indicator function. We also assume that $\varepsilon_i \sim$ i.i.d. N(0,1), $i = 1, \dots, n$. In this setup, at each allocation step the investigator attempts to intelligently guess the upcoming treatment assignment and selectively enroll a patient who, in their view, would be most suitable for the upcoming treatment. The investigator's strategy is to guess the treatment as one that has been less frequently assigned thus far, or make a random guess in case the current treatment numbers are equal. Assuming that the investigator favors an experimental treatment and is interested in demonstrating its superiority over the control, the biasing mechanism is as follows: at the $(i+1)$st step, a "healthier" subject is enrolled, if $D(i) < 0$ ($u_{i+1} = 0.5$); a "sicker" subject is enrolled, if $D(i) > 0$ ($u_{i+1} = -0.5$); or a "regular" subject is enrolled, if $D(i) = 0$ ($u_{i+1} = 0$).

We consider three statistical test procedures:

- ***T1: Two-sample t-test*:** The test statistic is $t = \frac{\bar{Y}_E - \bar{Y}_C}{\sqrt{S_p^2\left(\frac{1}{N_E(n)} + \frac{1}{N_C(n)}\right)}}$, where $\bar{Y}_E = \frac{1}{N_E(n)}\sum_{i=1}^{n} \delta_i Y_i$ and $\bar{Y}_C = \frac{1}{N_C(n)}\sum_{i=1}^{n}(1-\delta_i)Y_i$ are the treatment sample means, $N_E(n) = \sum_{i=1}^{n}\delta_i$ and $N_C(n) = n - N_E(n)$ are the observed group sample sizes, and S_p^2 is a pooled estimate of variance, where $S_p^2 = \frac{1}{n-2}\left(\sum_{i=1}^{n}\delta_i(Y_i - \bar{Y}_E)^2 + \sum_{i=1}^{n}(1-\delta_i)(Y_i - \bar{Y}_C)^2\right)$. Then $H_0: \Delta = 0$ is rejected at level α, if $|t| > t_{1-\frac{\alpha}{2}, n-2}$, the $100(1-\frac{\alpha}{2})$th percentile of the t-distribution with $n-2$ degrees of freedom.
- ***T2: Randomization-based test using mean difference*:** Let $\boldsymbol{\delta}_{obs}$ and $\boldsymbol{y}_{obs}$ denote, respectively the observed sequence of treatment assignments and responses, obtained from the trial using randomization procedure $\Re$. We first compute the observed mean difference

$S_{obs} = S(\boldsymbol{\delta}_{obs}, \boldsymbol{y}_{obs}) = \bar{Y}_E - \bar{Y}_C$. Then we use Monte Carlo simulation to generate L randomization sequences of length n using procedure $\Re$, where L is some large number. For the ℓth generated sequence, $\boldsymbol{\delta}_\ell$, compute $S_\ell = S(\boldsymbol{\delta}_\ell, \boldsymbol{y}_{obs})$, $\ell = 1, \dots, L$. Then the proportion of sequences for which S_ℓ is at least as extreme as S_{obs} is a computed as $\hat{P} = \frac{1}{L}\sum_{\ell=1}^{L} 1\{|S_\ell| \geq |S_{obs}|\}$. Statistical significance is declared, if $\hat{P} < \alpha$.

- ***T3: Randomization-based test based on ranks*:** This test procedure follows the same logic as T2, except that the test statistic is calculated based on ranks. Given the vector of observed responses $\boldsymbol{y}_{obs} = (y_1, \dots, y_n)$, let a_{jn} denote the rank of y_j among the elements of $\boldsymbol{y}_{obs}$. Let $\bar{a}_n$ denote the average of a_{jn}'s, and let $\boldsymbol{a}_n = (a_{1n} - \bar{a}_n, \dots, a_{nn} - \bar{a}_n)'$. Then a linear rank test statistic has the form $S_{obs} = \boldsymbol{\delta}'_{obs}\boldsymbol{a}_n = \sum_{i=1}^{n} \delta_i(a_{in} - \bar{a}_n)$.

We consider four scenarios of the true mean difference $\Delta = \mu_E - \mu_C$, which correspond to the Null case ($\Delta = 0$), and three choices of $\Delta > 0$ which correspond to Alternative 1 (power ~70%), Alternative 2 (power ~80%) , and Alternative 3 (power ~90%). In all cases, $n = 50$ was used.

Figure 4 summarizes the results of a simulation study comparing 11 randomization designs, under 4 models for the outcome (M1, M2, M3, and M4), 4 scenarios for the mean treatment difference (Null, and Alternatives 1, 2, and 3), using 3 statistical tests (T1, T2, and T3). The operating characteristics of interest are the type I error rate under the Null scenario and the power under the Alternative scenarios.

From Figure 4, under the normal random sampling model (M1), all considered randomization designs have similar performance: they maintain the type I error rate and have similar power, with all tests. In other words, when population model assumptions are satisfied, any combination of design and analysis should work well and yield reliable and consistent results.

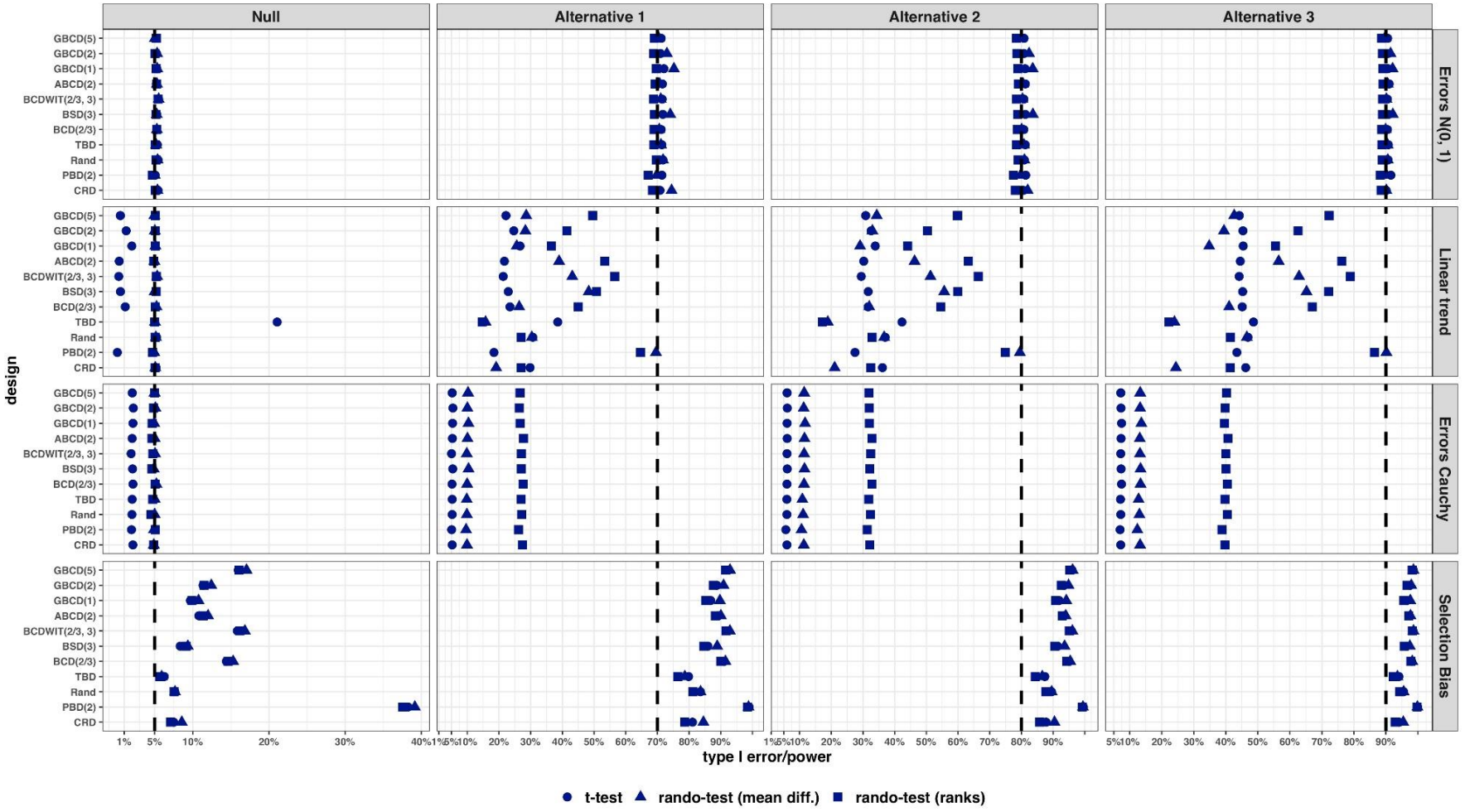
Under the "linear trend" model (M2), the designs have differential performance. First of all, under the Null scenario, only Rand and CRD maintain the type I error rate at 5% with all three tests. For TBD, the t-test is anticonservative, with type I error rate ~20%, whereas for eight other procedures the t-test is conservative, with type I error rate in the range 0.1% – 2%. At the same time, for all 11 designs the two randomization-based tests maintain the nominal type I error rate at 5%. These results are consistent with some previous findings in the literature [67, 68]. As regards power, it is reduced significantly compared to the normal random sampling scenario. The t-test

seems to be most affected and the randomization-based test using ranks is most robust for a majority of the designs (except for CRD, for which the t-test is the most powerful among the three tests). This signifies the usefulness of randomization-based inference in situations when outcome data are subject to a linear time trend, and the importance of applying randomization-based tests at least as supplemental analyses to likelihood-based test procedures.

Under the "Cauchy errors" model (M3), all designs perform similarly: the randomization-based tests maintain the type I error rate at 5%, whereas the t-test deflates the type I error to 2%. As regards power, all designs also have similar and consistent performance: the t-test is least powerful, and the randomization-based test using ranks has highest power among the three tests. Overall, under misspecification of the error distribution a randomization-based test using ranks is most appropriate; yet one should acknowledge that its power is still lower than expected.

Under the "selection bias" model (M4), the 11 designs have differential performance. First of all, there is evidence of some inflation of the type I error even for CRD (~7%), with all three tests. This is not a numerical artifact: we ran additional simulations under M4 (results are not shown here) and found that inflation of the type I error depends on the magnitude of the bias effect, but not on the sample size n. Second, from Figure 4, the type I error rate was a bit inflated for Rand (~7%), and to a lesser extent for TBD (~6%), with all three tests. This is consistent with the theory of Blackwell and Hodges [28] which posits that TBD provides best protection against selection bias within a class of restricted randomization procedures that force exact balance. For eight other procedures, greater inflations of the type I error were observed. As expected, the most affected design was PBD(2) for which the type I error rate was 38% – 39%. In general, the more random the design, the less susceptible it is to selection bias. For instance, BSD(3), which was among top three performing designs in terms of balance/randomness tradeoff (cf. Table 2) had 8% – 9% type I error rate with the three tests. Finally, under M4, power is inflated by several percentage points compared to the normal random sampling scenario without selection bias.

Figure 4: Simulated type I error rate and power of 11 restricted randomization procedures. Four models for the data generating mechanism of the primary outcome (M1: Normal random sampling; M2: Linear trend; M3: Errors Cauchy; and M4: Selection bias). Four scenarios for the treatment mean difference (Null; Alternatives 1, 2, and 3). Three statistical tests (T1: two-sample t-test; T2: randomization-based test using mean difference; T3: randomization-based test using ranks)



How can we reduce predictability of a randomization design and lower the risk of selection bias?

Selection bias can arise if the investigator can intelligently guess at least part of the randomization sequence yet to be allocated and, on that basis, preferentially and strategically assigns study subjects to treatments. Although it is generally not possible to prove that a particular study has been infected with selection bias, there are examples of published RCTs that do show some evidence to have been affected by it. Suspect trials are, for example, those with strong observed baseline covariate imbalances that consistently favor the active treatment group [16].

Consider the RCT of etanercept in children with polyarticular juvenile rheumatoid arthritis [82]. This trial had a three-month open-label run-in on etanercept, after which the treatment responders entered a double-blind period and were randomized to either placebo or etanercept for 4 months or until a flare of the disease occurred. Stratified permuted block randomization was used with blocks of size two. Stratification was based on study center and the number of active joints (≤ 2 vs. > 2). The trial concluded with a highly significant ($p<0.001$) treatment effect of etanercept vs. placebo. However, there was also evidence of group differences at baseline: patients in the etanercept group were overall younger and on different background medication, which could be indicative of a different level of disease severity.

In the etanercept trial, the randomization was, as noted, implemented in blocks of two. This block size should never be used because the risk of unmasking is unacceptable: every other patient is known to be on the opposite treatment from the one that was randomized immediately prior. This trial however, goes a step further: an open-label run-in period took place prior to randomization, during which the investigator would have been able to observe the specifics of the treatment effect. With this background knowledge, the actual treatment assignments (and therefore the subsequent treatment assignments) could have easily been deduced. Another problem was that the two prognostic strata (active joints ≤ 2 vs. > 2) within a center used blocks that were mirror images of each other, meaning that if one block was EC, then its matching block was necessarily CE. This means that unmasking one allocation would lead to perfect knowledge of three other allocations within the center.

Although we cannot with certainty conclude that it was selection bias that led this trial to be a successful trial, we can also not conclude that selection bias did in fact not impact this trial's results. What is disturbing is that by simply changing the randomization methodology from

permuted blocks to a less restrictive procedure that is not as conducive to bias, the trial would have had a greater level of randomization encryption and the results would have carried greater credibility. To illustrate the latter point, let us compare predictability of two randomization procedures – permuted block design (PBD) and big stick design (BSD) for several values of the maximum tolerated imbalance (MTI).

Table 3 reports two metrics for PBD and BSD: proportion of deterministic assignments within a randomization sequence, and excess correct guess probability. The latter metric is the absolute increase in proportion of correct guesses for a given procedure over CRD that has 50% probability of correct guesses under the "optimal guessing strategy"[1]. Note that for MTI=1, BSD is equivalent to PBD with blocks of two. However, by increasing MTI, one can substantially decrease predictability. For instance, going from MTI=1 in the BSD to an MTI of 2 or 3 (two bottom rows), the proportion of deterministic assignments decreases from 50% to 25% and 16.7%, respectively, and excess correct guess probability decreases from 25% to 12.5% and 8.3%, which is a substantial reduction in risk of selection bias. In addition to simplicity and lower predictability for the same level of MTI control, BSD has another important advantage: investigators are not accustomed to it (as they are to the PBD), and therefore it has potential for complete elimination of prediction through thwarting enough early prediction attempts.

Table 3: Predictability of permuted block design (PBD) and big stick design (BSD) for different values of maximum tolerated imbalance (MTI)

	Proportion of Deterministic Assignments		Excess Correct Guess Probability	
MTI	**PBD**	**BSD**	**PBD**	**BSD**
1	50%	50%	25%	25%
2	33.3%	25%	20.8%	12.5%
3	25%	16.7%	18.3%	8.3%

Our observations here are also generalizable to other MTI randomization methods, such as the maximal procedure [35], Chen's designs [38, 39], block urn design [40], just to name a few. MTI randomization procedures can be also used as building blocks for more complex stratified randomization schemes [83].

[1] Guess the next allocation as the treatment with fewest allocations in the sequence thus far, or make a random guess if the treatment numbers are equal.

How can we mitigate risk of chronological bias?

Chronological bias may occur if a trial recruitment period is long, and there is a drift in some covariate over time that is subsequently not accounted for in the analysis [27]. To mitigate risk of chronological bias, treatment assignments should be balanced over time. In this regard, the ICH E9 guideline has the following statement [31]:

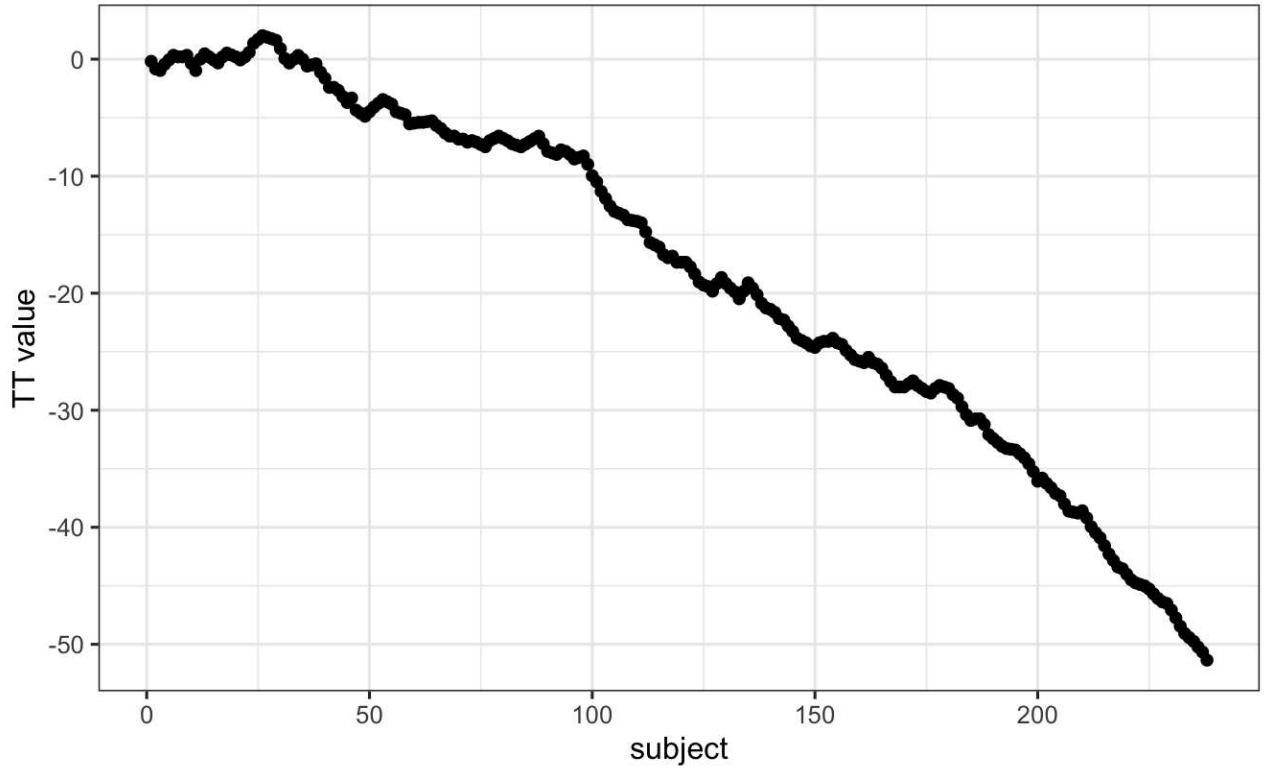
"...Although unrestricted randomisation is an acceptable approach, some advantages can generally be gained by randomising subjects in blocks. This helps to increase the comparability of the treatment groups, particularly when subject characteristics may change over time, as a result, for example, of changes in recruitment policy. It also provides a better guarantee that the treatment groups will be of nearly equal size..."

While randomization in blocks of two ensures best balance, it is highly predictable. In practice, a sensible tradeoff between balance and randomness is desirable. In the following example, we illustrate the issue of chronological bias in the context of a real RCT.

Altman and Royston [84] gave several examples of clinical studies with hidden time trends. For instance, an RCT to compare azathioprine versus placebo in patients with primary biliary cirrhosis (PBC) with respect to overall survival was an international, double-blind, randomized trial including 248 patients of whom 127 received azathioprine and 121 placebo [85]. The study had a recruitment period of 7 years. A major prognostic factor for survival was the serum bilirubin level on entry to the trial. Altman and Royston [84] provided a cusum plot of log bilirubin which showed a strong decreasing trend over time – patients who entered the trial later had, on average, lower bilirubin levels, and therefore better prognosis. Despite that the trial was randomized, there was some evidence of baseline imbalance with respect to serum bilirubin between azathioprine and placebo groups. The analysis using Cox regression adjusted for serum bilirubin showed that the treatment effect of azathioprine was statistically significant (p=0.01), with azathioprine reducing the risk of dying to 59% of that observed during the placebo treatment.

The azathioprine trial [85] provides a very good example for illustrating importance of both the choice of a randomization design and a subsequent statistical analysis. We evaluated several randomization designs and analysis strategies under the given time trend through simulation. Since we did not have access to the patient level data from the azathioprine trial, we simulated a dataset of serum bilirubin values from 248 patients that resembled that in the original paper (Figure 1 in [84]); see Figure 5 below.

Figure 5: Cusum plot of baseline log serum bilirubin level of 248 subjects from the azathioprine trial, reproduced from Figure 1 of Altman and Royston (1988)



For the survival outcomes, we use the following data generating mechanism [86, 87]: let $h_i(t, \delta_i)$ denote the hazard function of the ith patient at time t such that

$$h_i(t, \delta_i) = h_c(t)\exp(\delta_i \log HR + u_i),\ i = 1, \ldots, 248, \quad (2)$$

where $h_c(t)$ is an unspecified baseline hazard, $\log HR$ is the true value of the log-transformed hazard ratio, and u_i is the log serum bilirubin of the ith patient at study entry.

We considered 4 randomization designs: CRD, Rand, TBD, and PBD with blocks of two. To evaluate both type I error and power, we considered two values for the true treatment effect: $HR = 1$ (Null) and $HR = 0.6$ (Alternative). For data analysis, we used the Cox regression model, either with or without adjustment for serum bilirubin. For the sake of simplicity, we let $h_c(t) \equiv 1$ (exponential distribution) and assume no censoring when simulating the data.

For each combination of the design, experimental scenario, and data analysis strategy, a trial with 248 patients was simulated 10,000 times. In each simulation, we used the same time trend in serum bilirubin as described. Through simulation, we estimated the probability of a statistically

significant baseline imbalance in serum bilirubin between azathioprine and placebo groups, type I error rate, and statistical power.

First, we observed that the designs differ with respect to their potential to achieve baseline covariate balance under the time trend. For instance, probability of a statistically significant group difference on serum bilirubin (two-sided $P < 0.05$) is ~24% for TBD, ~10% for CRD, ~0.9% for Rand, and ~0% for PBD(2).

Second, a failure to adjust for serum bilirubin in the analysis can negatively impact statistical inference. Table 4 shows the type I error and power of statistical analyses unadjusted and adjusted for serum bilirubin. For the unadjusted analysis, only CR and Rand are valid (maintain the type I error rate at 5%), whereas TBD is anticonservative (~15% type I error) and PBD(2) is conservative (<1% type I error). These findings are consistent with the ones for the two-sample t-test described earlier in the current paper, and they agree well with other findings in the literature [67]. By contrast, the covariate-adjusted analysis is valid for all four randomization designs. Furthermore, for each design, unadjusted analysis is substantially less powerful (~59-65% power) than the corresponding covariate-adjusted analysis (~97% power). This speaks to the importance of covariate-adjusted analysis, which should be straightforward if a covariate affected by a time trend is known (e.g. serum bilirubin in our example). If a covariate is unknown or hidden, then unadjusted analysis may have reduced power and distorted type I error, but the designs such as CRD and Rand do ensure valid statistical inference. Alternatively, randomization-based tests can be applied. The resulting analysis will be valid but may be less powerful.

Table 4: Type I error and power of four randomization designs under a time trend

	Type I error rate		Power	
	Analysis unadjusted for serum bilirubin	Analysis adjusted for serum bilirubin	Analysis unadjusted for serum bilirubin	Analysis adjusted for serum bilirubin
CR	0.0481	0.0504	0.6114	0.9694
Rand	0.0517	0.0511	0.6193	0.9704
TBD	0.1451	0.0511	0.5856	0.9702
PBD(2)	0.0064	0.0511	0.6540	0.9700

How do we design an RCT with a very small sample size?

In our last example, we illustrate the importance of careful consideration for the choice of randomization design and subsequent statistical analysis in a nonstandard RCT with small sample size. Due to confidentiality and because this study is still in conduct, we do not disclose all details here except for that the study is an ongoing phase II RCT in a very rare and devastating autoimmune disease in children.

The study includes three periods: an open-label single-arm active treatment for 28 weeks to identify treatment responders (Period 1), a 24-week randomized treatment withdrawal period to primarily assess the efficacy of the active treatment vs. placebo (Period 2), and a 3-year long-term safety, open-label active treatment (Period 3). Because of a challenging indication and the rarity of the disease, the study plans to enroll up to 10 male or female pediatric patients (≤17 years of age) in order to randomize 8 patients (4 per treatment arm) in Period 2 of the study. The primary endpoint for assessing the efficacy of active treatment versus placebo is the proportion of patients with disease flare during the 24-week randomized withdrawal phase. The two groups will be compared using the Fisher's exact test. In case of a successful outcome, evidence of clinical efficacy from this study will be also used as part of a package to support the claim for drug effectiveness.

Very small sample sizes are not uncommon in clinical trials of rare diseases [88, 89]. Naturally, there are several methodological challenges for this type of study. A major challenge is generalizability of the results from the RCT to a population. In this particular indication, no approved treatment exists, and there is uncertainty on disease epidemiology and the exact number of patients with the disease who would benefit from treatment (patient horizon). Another challenge is the choice of the randomization procedure and the primary statistical analysis. In this study, one can enumerate upfront all 25 possible outcomes: {0, 1, 2, 3, 4} responders on active treatment, and {0, 1, 2, 3, 4} responders on placebo, and create a chart quantifying the level of evidence (p-value) for each experimental outcome, and the corresponding decision. Before the trial starts, a discussion with the regulatory agency is warranted to agree upon on what level of evidence must be achieved in order to declare the study a "success".

Let us perform a hypothetical planning for the given study. Suppose we go with a standard population-based approach, for which we test the hypothesis $H_0: p_E = p_C$ vs. $H_0: p_E > p_C$ (where

p_E and p_C stand for the true success rates for the experimental and control group, respectively) using Fisher's exact test. Table 5 provides 1-sided p-values of all possible experimental outcomes. One could argue that a p-value < 0.1 may be viewed as a convincing level of evidence for this study. There are only 3 possibilities that can lead to this outcome: 3/4 vs. 0/4 successes (p = 0.0714); 4/4 vs. 0/4 successes (p = 0.0143); and 4/4 vs. 1/4 successes (p = 0.0714). For all other outcomes, p ≥ 0.2143, and thus the study would be regarded as a "failure".

Table 5: All possible outcomes, p-values, and corresponding decisions for an RCT with n=8 patients (4 per treatment arm), for two randomization designs (Rand and TBD)

Number of responders		Difference in proportions (Experimental vs. Control)	Fisher's exact test 1-sided p-value	Decision*
Experimental	Control			
0/4	0/4	0	1.0	F
1/4	1/4	0	0.7857	F
2/4	2/4	0	0.7571	F
3/4	3/4	0	0.7857	F
4/4	4/4	0	1.0	F
1/4	0/4	0.25	0.5	F
2/4	0/4	0.50	0.2143	F
3/4	0/4	0.75	0.0714	S
4/4	0/4	1	0.0143	S
0/4	1/4	-0.25	1.0	F
0/4	2/4	-0.50	1.0	F
0/4	3/4	-0.75	1.0	F
0/4	4/4	-1	1.0	F
2/4	1/4	0.25	0.5	F
3/4	1/4	0.50	0.2429	F
4/4	1/4	0.75	0.0714	S
1/4	2/4	-0.25	0.9286	F
1/4	3/4	-0.50	0.9857	F
1/4	4/4	-0.75	1.0	F
3/4	2/4	0.25	0.5	F
4/4	2/4	0.50	0.2143	F
2/4	3/4	-0.25	0.9286	F
2/4	4/4	-0.50	1.0	F
4/4	3/4	0.25	0.5	F
3/4	4/4	-0.25	1.0	F

* F = Declare study a failure; S = Declare study a success

Now let us consider a randomization-based inference approach. For illustration purposes, we consider two restricted randomization procedures—Rand and TBD—that exactly achieve 4:4 allocation. The reference set for a randomization design with 4:4 allocation includes $70 = \binom{8}{4}$ unique sequences though with different probabilities of observing each sequence. For Rand, these sequences are equiprobable, whereas for TBD, some sequences are more likely than others. In practice, the study statistician picks a treatment sequence at random from the reference set according to the chosen design. The details (randomization seed, chosen sequence, etc.) are carefully documented and kept confidential. For the chosen sequence and the observed outcome data, a randomization-based p-value is the sum of probabilities of all sequences in the reference set that yield the result at least as large in favor of the experimental treatment as the one observed. This p-value will depend on the randomization design, the observed randomization sequence and the observed outcomes, and it may also be different from the population-based analysis p-value.

To illustrate this, suppose the chosen randomization sequence is CEECEECC (C stands for control and E stands for experimental), and the observed responses are FSSFSFFF (F stands for failure and S stands for success). Thus, we have 3/4 successes on experimental and 0/4 successes on control. Then, the randomization-based p-value is 0.0714 for Rand; it is 0.0938 for TBD; and it is 0.0714 for the population-based analysis. The coincidence of the randomization-based p-value for Rand and the p-value of the population-based analysis is not surprising. Fisher's exact test is a permutation test and in the case of Rand as randomization procedure the p-value of a permutation test and of a randomization test are always equal. However, despite the numerical equality, we should be mindful of different assumptions (population / randomization model).

Likewise, randomization-based p-values can be derived for other combinations of observed randomization sequences and responses. All these details (the chosen randomization design, the analysis strategy, and corresponding decisions) would have to be fully specified upfront (before the trial starts) and agreed upon by both the sponsor and the regulator. This would remove any ambiguity when the trial data become available.

Conclusions

Summary and discussion

Randomization is the foundation of any RCT involving treatment comparison. Randomization is not a single technique, but a very broad class of statistical methodologies for design and analysis

of clinical trials [9]. In this paper, we focused on the randomized controlled two-arm trial designed with equal allocation, which is the gold standard research design to generate clinical evidence in support of regulatory submissions. Even in this relatively simple case, there are various restricted randomization procedures with different probabilistic structures and different statistical properties.

For the 1:1 RCT, there is a dual goal of balancing treatment assignments while maintaining allocation randomness. Final balance in treatment totals frequently maximizes statistical power for treatment comparison. It is also important to maintain balance at intermediate steps during the trial, especially in long-term studies, to mitigate potential for chronological bias. At the same time, a procedure should have high degree of randomness so that treatment assignments within the sequence are not easily predictable; otherwise, the procedure may be vulnerable to selection bias, especially in open-label studies. While balance and randomness are competing criteria, it is possible to find restricted randomization procedures that provide a sensible tradeoff between these criteria, e.g. the MTI procedures, of which the big stick design [37] with a suitably chosen MTI limit has very appealing statistical properties. In practice, the choice of a randomization procedure should be made after a systematic evaluation of different candidate procedures under different experimental scenarios for the primary outcome, including cases when model assumptions are violated.

In our considered examples we showed that the choice of randomization design, data analytic technique (e.g. parametric or nonparametric model), and the decision on whether to include randomization in the analysis (e.g. randomization-based or population model-based analysis) are all very important considerations. Furthermore, these examples highlight the importance of using randomization designs that provide strong encryption of the randomization sequence, importance of covariate adjustment in the analysis, and the value of statistical thinking in nonstandard RCTs with very small sample sizes and small patient horizon. Finally, in this paper we have discussed randomization-based tests as robust and valid alternatives to likelihood-based tests. Randomization-based inference is a useful approach in clinical trials and should be considered more broadly by clinical researchers [14].

Further topics on randomization

Given the breadth of the subject of randomization, many important topics have been omitted from the current paper. Here we outline just a few of them.

In this paper, we have focused on the 1:1 RCT. However, clinical trials may involve more than two treatment arms. Extensions of equal randomization to the case of multiple treatment arms is relatively straightforward for many restricted randomization procedures [9]. Some trials with two or more treatment arms use unequal allocation (e.g. 2:1). Randomization procedures with unequal allocation ratios require careful consideration. For instance, an important and desirable feature is the allocation ratio preserving property (ARP). A randomization procedure targeting unequal allocation is said to be ARP, if at each allocation step the unconditional probability of a particular treatment assignment is the same as the target allocation proportion for this treatment [90]. Non-ARP procedures may have fluctuations in the unconditional randomization probability from allocation to allocation, which may be problematic [91]. Fortunately, some randomization procedures naturally possess the ARP property, and there are approaches to correct for a non-ARP deficiency – these should be considered in the design of RCTs with unequal allocation ratios [90, 91, 92].

In many RCTs, investigators may wish to prospectively balance treatment assignments with respect to important prognostic covariates. For a small number of categorical covariates one can use stratified randomization by applying separate MTI randomization procedures within strata [83]. However, a potential advantage of stratified randomization decreases as the number of stratification variables increases [93]. In trials where balance over a large number of covariates is sought and the sample size is small or moderate, one can consider covariate-adaptive randomization procedures that achieve balance within covariate margins, such as the minimization procedure [94, 95], optimal model-based procedures [46], or some other covariate-adaptive randomization technique [96]. To achieve valid and powerful results, covariate-adaptive randomization design must be followed by covariate-adjusted analysis [97]. Special considerations are required for covariate-adaptive randomization designs with more than two treatment arms and/or unequal allocation ratios [98].

In some clinical research settings, such as trials for rare and/or life threatening diseases, there is a strong ethical imperative to increase the chance of a trial participant to receive an empirically better treatment. Response-adaptive randomization (RAR) has been increasingly considered in practice, especially in oncology [99, 100]. Very extensive methodological research on RAR has been done [101, 102]. RAR is increasingly viewed as an important ingredient of complex clinical trials such as umbrella and platform trial designs [103, 104]. While RAR, when properly applied,

has its merit, the topic has generated a lot of controversial discussions over the years [105–109]. Amid the ongoing COVID-19 pandemic, RCTs evaluating various experimental treatments for critically ill COVID-19 patients do incorporate RAR in their design; see, for example, the I-SPY COVID-19 trial (https://clinicaltrials.gov/ct2/show/NCT04488081).

Randomization can also be applied more broadly than in conventional RCT settings where randomization units are individual subjects. For instance, in a cluster randomized trial, not individuals but groups of individuals (clusters) are randomized among one or more interventions or the control [110]. Observations from individuals within a given cluster cannot be regarded as independent, and special statistical techniques are required to design and analyze cluster-randomized experiments. In some clinical trial designs, randomization is applied within subjects. For instance, the micro-randomized trial (MRT) is a novel design for development of mobile treatment interventions in which randomization is applied to select different treatment options for individual participants over time to optimally support individuals' health behaviors [111].

Finally, beyond the scope of the present paper are the regulatory perspectives on randomization and practical implementation aspects, including statistical software and information systems to generate randomization schedules in real time. We hope to cover these topics in subsequent papers.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All results reported in this paper are based either on theoretical considerations or simulation evidence. The computer code (using R and Julia programming languages) is fully documented and is available upon reasonable request.

Competing interests

None.

Funding

None. The opinions expressed in this article are those of the authors and may not reflect the opinions of the organizations that they work for.

Authors' contributions

Conception: VWB, KC, NH, RDH, OS. Writing of the main manuscript: OS, with contributions from VWB, KC, JJC, CE, NH, and RDH. Design of simulation studies: OS, YR. Development of code and running simulations: YR. Digitization and preparation of data for Figure 5: JR. All authors reviewed the manuscript.

Acknowledgement

The authors are grateful to Robert A. Beckman for his continuous efforts coordinating Innovative Design Scientific Working Groups, which is also a networking research platform for the Randomization ID SWG.

References

1. Byar DP, Simon RM, Friedewald WT, Schlesselman JJ, DeMets DL, Ellenberg JH, Gail MH, Ware JH. Randomized clinical trials—perspectives on some recent ideas. The New England Journal of Medicine. 1976; 295:74-80.
2. Collins R, Bowman L, Landray M, Peto R. The magic of randomization versus the myth of real-world evidence. The New England Journal of Medicine. 2020; 382:674-678.
3. ICH Harmonised Tripartite Guideline. General Considerations for Clinical Trials E8. 1997.
4. Byar DP. Why data bases should not replace randomized clinical trials. Biometrics. 1980; 36:337-342.
5. Mehra MR, Desai SS, Kuy SR, Henry TD, Patel AN. Cardiovascular disease, drug therapy, and mortality in Covid-19. The New England Journal of Medicine. 2020; 382:e102. DOI: 10.1056/NEJMoa2007621 (RETRACTED PAPER).
6. Mehra MR, Desai SS, Ruschitzka F, Patel AN. Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. The Lancet. 2020, https://doi.org/10.1016/S0140-6736(20)31180-6 (RETRACTED PAPER).
7. Mehra MR, Desai SS, Kuy SR, Henry TD, Patel AN. Retraction: Cardiovascular disease, drug therapy, and mortality in Covid-19. The New England Journal of Medicine. 2020. DOI: 10.1056/NEJMoa2007621.
8. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. British Medical Journal. 1948; 2:769-782.
9. Rosenberger WF, Lachin J. Randomization in clinical trials: theory and practice. 2nd ed. Wiley, New York; 2015.

10. Fisher RA. The Design of Experiments. Edinburgh: Oliver and Boyd; 1935.
11. Hill AB. The clinical trial. British Medical Bulletin. 1951; 7(4):278-282.
12. Hill AB. Memories of the British streptomycin trial in tuberculosis: The first randomized clinical trial. Controlled Clinical Trials. 1990; 11:77-79.
13. Armitage PA. The role of randomization in clinical trials. Statistics in Medicine. 1982; 1:345-352.
14. Rosenberger WF, Uschner D, Wang Y. Randomization: The forgotten component of the randomized clinical trial. Statistics in Medicine. 2019; 38(1):1-30 (with discussion).
15. Berger VW. Trials: the worst possible design (except for all the rest). The International Journal of Person Centered Medicine. 2011; 1(3):630-631.
16. Berger VW. Selection bias and covariate imbalances in randomized clinical trials. Wiley, New York; 2005.
17. Berger VW. The alleged benefits of unrestricted randomization. In: Berger VW, editor. Randomization, masking, and allocation concealment. CRC Press, Boca Raton FL; 2018. p.39-50.
18. Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? British Medical Journal. 1999; 318:1209.
19. Senn S. Testing for baseline balance in clinical trials. Statistics in Medicine. 1994; 13:1715-1726.
20. Senn S. Seven myths of randomisation in clinical trials. Statistics in Medicine. 2013; 32:1439-1450.
21. Rosenberger WF, Sverdlov O. Handling covariates in the design of clinical trials. Statistical Science. 2008; 23:404-419.
22. Proschan M, Dodd L. Re-randomization tests in clinical trials. Statistics in Medicine. 2019; 38:2292-2302.
23. Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. Journal of the Royal Statistical Society: Series A (Statistics in Society). 1994 May; 157(3):357-387.
24. Berry SM, Carlin BP, Lee JJ, Muller P. Bayesian adaptive methods for clinical trials. CRC Press, Boca Raton FL; 2010.
25. Lachin J. Properties of simple randomization in clinical trials. Controlled Clinical Trials. 1988; 9:312-326.
26. Pocock SJ. Allocation of patients to treatment in clinical trials. Biometrics. 1979; 35(1):183-197.
27. Simon R. Restricted randomization designs in clinical trials. Biometrics. 1979; 35(2):503-512.
28. Blackwell D, Hodges JL. Design for the control of selection bias. The Annals of Mathematical Statistics. 1957; 28(2):449-460.

29. Matts JP, McHugh R. Analysis of accrual randomized clinical trials with balanced groups in strata. Journal of Chronic Diseases. 1978; 31:725-740.

30. Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. Controlled Clinical Trials. 1988; 9:327-344.

31. ICH Harmonised Tripartite Guideline. Statistical Principles for Clinical Trials E9. 1998.

32. Shao H, Rosenberger WF. Properties of the random block design for clinical trials. In: Kunert J, Müller CH, Atkinson AC, eds. mODa 11 – Advances in model-oriented design and analysis. Springer International Publishing Switzerland; 2016. p.225-233.

33. Zhao W. Evolution of restricted randomization with maximum tolerated imbalance. In: Berger VW, editor. Randomization, masking, and allocation concealment. CRC Press, Boca Raton FL; 2018. p.61-81.

34. Bailey RA, Nelson PR. Hadamard randomization: a valid restriction of random permuted blocks. Biometrical Journal. 2003; 45(5):554-560.

35. Berger VW, Ivanova A, Knoll MD. Minimizing predictability while retaining balance through the use of less restrictive randomization procedures. Statistics in Medicine. 2003; 22:3017-3028.

36. Zhao W, Berger VW, Yu Z. The asymptotic maximal procedure for subject randomization in clinical trials. Statistical Methods in Medical Research. 2018; 27(7):2142-2153.

37. Soares JF, Wu CFJ. Some restricted randomization rules in sequential designs. Communications in Statistics—Theory and Methods. 1983; 12(17):2017-2034.

38. Chen YP. Biased coin design with imbalance tolerance. Communications in Statistics. Stochastic Models. 1999; 15(5):953-975.

39. Chen YP. Which design is better? Ehrenfest urn versus biased coin. Advances in Applied Probability. 2000; 32:738-749.

40. Zhao W, Weng Y. Block urn design—A new randomization algorithm for sequential trials with two or more treatments and balanced or unbalanced allocation. Contemporary Clinical Trials. 2011; 32:953-961.

41. van der Pas SL. Merged block randomisation: A novel randomisation procedure for small clinical trials. Clinical Trials. 2019; 16(3):246-252.

42. Zhao W. Letter to the Editor – Selection bias, allocation concealment and randomization design in clinical trials. Contemporary Clinical Trials. 2013; 36:263-265.

43. Berger VW, Bejleri K, Agnor R. Comparing MTI randomization procedures to blocked randomization. Statistics in Medicine. 2016; 35:685-694.

44. Efron B. Forcing a sequential experiment to be balanced. Biometrika. 1971; 58(3):403-417.

45. Wei LJ. The adaptive biased coin design for sequential experiments. The Annals of Statistics. 1978; 6(1):92-100.

46. Atkinson AC. Optimum biased coin designs for sequential clinical trials with prognostic factors. Biometrika. 1982; 69(1):61-67.

47. Smith RL. Sequential treatment allocation using biased coin designs. Journal of the Royal Statistical Society, Series B. 1984; 46(3):519-543.

48. Ball FG, Smith AFM, Verdinelli I. Biased coin designs with a Bayesian bias. Journal of Statistical Planning and Inference. 1993; 34(3):403-421.

49. Baldi Antognini A, Giovagnoli A. A new 'biased coin design' for the sequential allocation of two treatments. Applied Statistics. 2004; 53(4):651-664.

50. Atkinson AC. Selecting a biased-coin design. Statistical Science. 2014; 29(1):144-163.

51. Rosenberger WF. Randomized urn models and sequential design. Sequential Analysis. 2002; 21(1&2): 1-41 (with discussion).

52. Wei LJ, Lachin JM. Properties of the urn randomization in clinical trials. Controlled Clinical Trials. 1988; 9:345-364.

53. Wei LJ. An application of an urn model to the design of sequential controlled clinical trials. Journal of the American Statistical Association. 1978; 73(363):559-563.

54. Schouten HJA. Adaptive biased urn randomization in small strata when blinding is impossible. Biometrics. 1995; 51(4):1529-1535.

55. Ivanova A. A play-the-winner-type urn design with reduced variability. Metrika. 2003; 58:1-13.

56. Kundt G. A new proposal for setting parameter values in restricted randomization methods. Methods of Information in Medicine. 2007; 46(4):440-449.

57. Kalish LA, Begg CB. Treatment allocation methods in clinical trials: a review. Statistics in Medicine. 1985; 4:129-144.

58. Zhao W, Weng Y, Wu Q, Palesch Y. Quantitative comparison of randomization designs in sequential clinical trials based on treatment balance and allocation randomness. Pharmaceutical Statistics. 2012; 11:39-48.

59. Flournoy N, Haines LM, Rosenberger WF. A graphical comparison of response-adaptive randomization procedures. Statistics in Biopharmaceutical Research. 2013; 5(2):126-141.

60. Hilgers RD, Uschner D, Rosenberger, WF, Heussen N. ERDO – a framework to select an appropriate randomization procedure for clinical trials. BMC Medical Research Methodology. 2017; 17:159.

61. Burman CF. On Sequential Treatment Allocations in Clinical Trials. PhD Thesis Dept. Mathematics, Göteborg. 1996.

62. Azriel D, Mandel M, Rinott Y. Optimal allocation to maximize the power of two-sample tests for binary response. Biometrika. 2012; 99(1):101-113.

63. Begg CB, Kalish LA. Treatment allocation for nonlinear models in clinical trials: the logistic model. Biometrics. 1984; 40:409-420.

64. Kalish LA, Harrington DP. Efficiency of balanced treatment allocation for survival analysis. Biometrics. 1988; 44(3):815-821.

65. Sverdlov O, Rosenberger WF. On recent advances in optimal allocation designs for clinical trials. Journal of Statistical Theory and Practice. 2013; 7(4):753-773.

66. Sverdlov O, Ryeznik Y, Wong WK. On optimal designs for clinical trials: an updated review. Journal of Statistical Theory and Practice. 2020; 14:10.

67. Rosenkranz GK. The impact of randomization on the analysis of clinical trials. Statistics in Medicine. 2011; 30:3475-3487.

68. Galbete A, Rosenberger WF. On the use of randomization tests following adaptive designs. Journal of Biopharmaceutical Statistics. 2016; 26(3):466-474.

69. Proschan M. Influence of selection bias on type I error rate under random permuted block design. Statistica Sinica. 1994; 4:219-231.

70. Kennes LN, Cramer E, Hilgers RD, Heussen N. The impact of selection bias on test decisions in randomized clinical trials. Statistics in Medicine. 2011; 30:2573-2581.

71. Rückbeil MV, Hilgers RD, Heussen N. Assessing the impact of selection bias on test decisions in trials with a time-to-event outcome. Statistics in Medicine. 2017; 36:2656-2668.

72. Uschner D, Hilgers RD, Heussen N. The impact of selection bias in randomized multi-arm parallel group clinical trials. PLoS ONE. 2018; 13(1):e0192065.

73. Berger VW, Exner DV. Detecting selection bias in randomized clinical trials. Controlled Clinical Trials. 1999; 25:515-524.

74. Ivanova A, Barrier RC, Berger VW. Adjusting for observable selection bias in block randomized trials. Statistics in Medicine. 2005; 24:1537-1546.

75. Kennes LN, Rosenberger WF, Hilgers RD. Inference for blocked randomization under a selection bias model. Biometrics. 2015; 71:979-984.

76. Hilgers RD, Manolov M, Heussen N, Rosenberger WF. Design and analysis of stratified clinical trials in the presence of bias. Statistical Methods in Medical Research. 2020; 29(6):1715-1727.

77. Zhao W. Letter to the Editor – A better alternative to the inferior permuted block design is not necessarily complex. Statistics in Medicine. 2016; 35:1736-1738.

78. Berger VW. Pros and cons of permutation tests in clinical trials. Statistics in Medicine. 2000; 19:1319-1328.

79. Simon R, Simon NR. Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization. Statistics and Probability Letters. 2011; 81:767-772.

80. Chow SC, Shao J, Wang H, Lokhnygina. Sample size calculations in clinical research. 3rd ed. CRC Press, Boca Raton FL; 2018.

81. Heritier S, Gebski V, Pillai A. Dynamic balancing randomization in controlled clinical trials. Statistics in Medicine. 2005; 24:3729-3741.

82. Lovell DJ, Giannini EH, Reiff A, et al. Etanercept in children with polyarticular juvenile rheumatoid arthritis. The New England Journal of Medicine. 2000; 342(11):763-769.

83. Zhao W. A better alternative to stratified permuted block design for subject randomization in clinical trials. Statistics in Medicine. 2014; 33:5239-5248.

84. Altman DG, Royston JP. The hidden effect of time. Statistics in Medicine. 1988; 7:629-637.

85. Christensen E, Neuberger J, Crowe J, et al. Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis. Gastroenterology. 1985; 89:1084-1091.

86. Rückbeil MV, Hilgers RD, Heussen N. Assessing the impact of selection bias on test decisions in trials with a time-to-event outcome. Statistics in Medicine. 2017; 36:2656-2668.

87. Rückbeil MV, Hilgers RD, Heussen N. Randomization in survival trials: An evaluation method that takes into account selection and chronological bias. PLoS One. 2019; 14(6):e0217964.

88. Hilgers RD, König F, Molenberghs G, Senn S. Design and analysis of clinical trials for small rare disease populations. Journal of Rare Diseases Research & Treatment 2016; 1(3):53-60.

89. Miller F, Zohar S, Stallard N, Madan J, Posch M, Hee SW, Pearce M, Vågerö M, Day S. Approaches to sample size calculation for clinical trials in rare diseases. Pharmaceutical Statistics. 2017; 17:214-230.

90. Kuznetsova OM, Tymofyeyev Y. Preserving the allocation ratio at every allocation with biased coin randomization and minimization in studies with unequal allocation. Statistics in Medicine. 2012; 31(8):701-723.

91. Kuznetsova OM, Tymofyeyev Y. Brick tunnel and wide brick tunnel randomization for studies with unequal allocation. In: Sverdlov O, editor. Modern adaptive randomized clinical trials: statistical and practical aspects. CRC Press, Boca Raton FL; 2015. p.83-114.

92. Kuznetsova OM, Tymofyeyev Y. Expansion of the modified Zelen's approach randomization and dynamic randomization with partial block supplies at the centers to unequal allocation. Contemporary Clinical Trials. 2011; 32:962-972.

93. EMA. Guideline on adjustment for baseline covariates in clinical trials. 26 February 2015.

94. Taves DR. Minimization: A new method of assigning patients to treatment and control groups. Clinical Pharmacology and Therapeutics. 1974; 15(5):443-453.

95. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics. 1975; 31(1):103-115.

96. Hu F, Hu Y, Ma Z, Rosenberger WF. Adaptive randomization for balancing over covariates. Wiley Interdisciplinary Reviews: Computational Statistics. 2014; 6(4):288-303.

97. Senn S. Statistical Issues in Drug Development. 2nd ed. Wiley-Interscience; 2007.

98. Kuznetsova OM, Tymofyeyev Y. Covariate-adaptive randomization with unequal allocation. In: Sverdlov O, editor. Modern adaptive randomized clinical trials: statistical and practical aspects. CRC Press, Boca Raton FL; 2015. p.171-197.

99. Berry DA. Adaptive clinical trials: the promise and the caution. Journal of Clinical Oncology. 2011; 29(6):606-609.

100. Trippa L, Lee EQ, Wen PY, Batchelor TT, Cloughesy T, Parmigiani G, Alexander BM. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. Journal of Clinical Oncology. 2012; 30(26):3258-3263.

101. Hu F, Rosenberger WF. The theory of response-adaptive randomization in clinical trials. Wiley, New York; 2006.

102. Atkinson AC, Biswas A. Randomised response-adaptive designs in clinical trials. CRC Press, Boca Raton FL; 2014.

103. Rugo HS, Olopade OI, DeMichele A, et al. Adaptive randomization of veliparib–carboplatin treatment in breast cancer. The New England Journal of Medicine. 2016; 375:23-34.

104. Berry SM, Petzold EA, Dull P, et al. A response-adaptive randomization platform trial for efficient evaluation of Ebola virus treatments: a model for pandemic response. Clinical Trials. 2016; 13:22-30.

105. Ware JH. Investigating therapies of potentially great benefit: ECMO. (with discussion). Statistical Science. 1989; 4(4):298-340.

106. Hey SP, Kimmelman J. Are outcome-adaptive allocation trials ethical? (with discussion). Clinical Trials. 2005; 12(2):102-127.

107. Proschan M, Evans S. The temptation of response-adaptive randomization. Clinical Infectious Diseases. 2020; ciaa334; doi: 10.1093/cid/ciaa334.

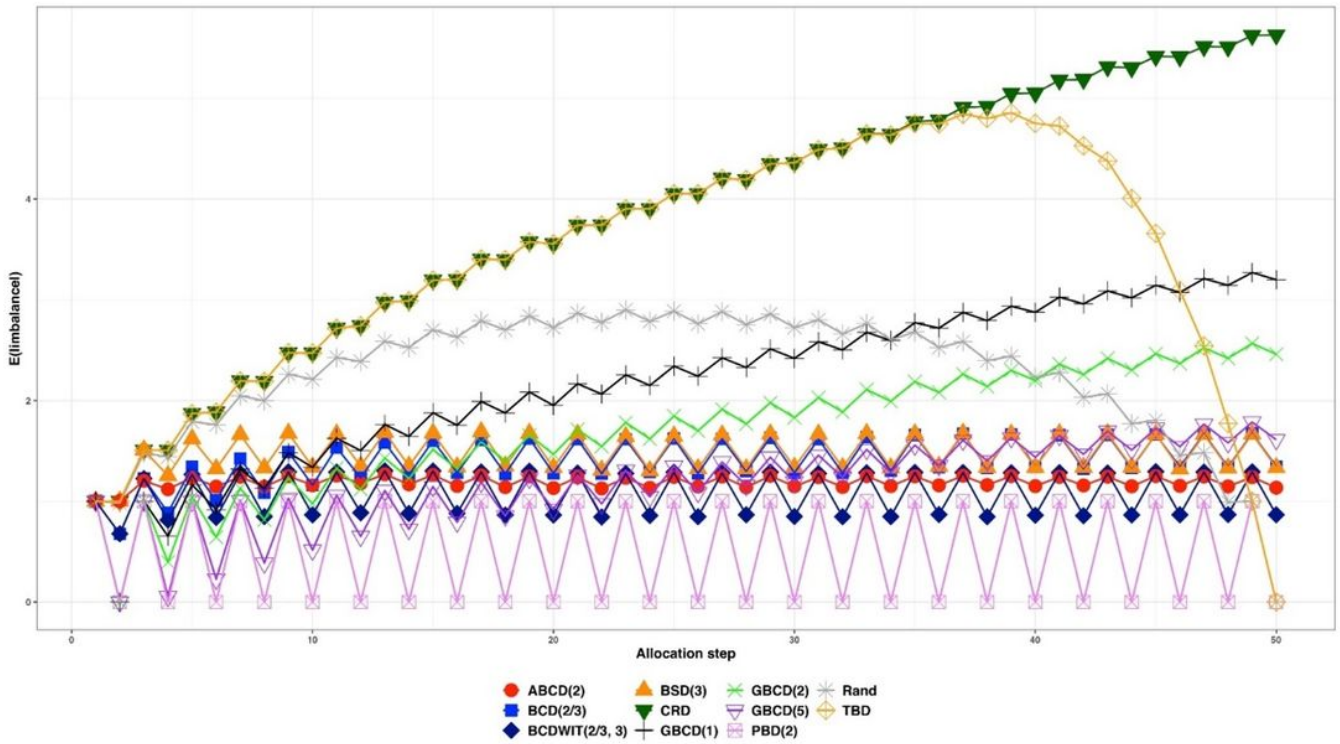
108. Villar SS, Robertson DS, Rosenberger WF. The temptation of overgeneralizing response-adaptive randomization. Clinical Infectious Diseases. 2020; ciaa1027; doi: 10.1093/cid/ciaa1027.

109. Proschan M. Reply to Villar, et al. *Clinical Infectious Diseases*. 2020; ciaa1029; doi: 10.1093/cid/ciaa1029.

110. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold Publishers Limited; 2000.

111. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, Murphy SA. Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions. Health Psychology. 2015; 34:1220-1228.

Figures



Note: PBD(2) has a forced periodicity absolute imbalance of 0, which distinguishes it from an MTI procedure

Figure 1

Simulated expected absolute imbalance vs. allocation step for 11 restricted randomization procedures for n=50.

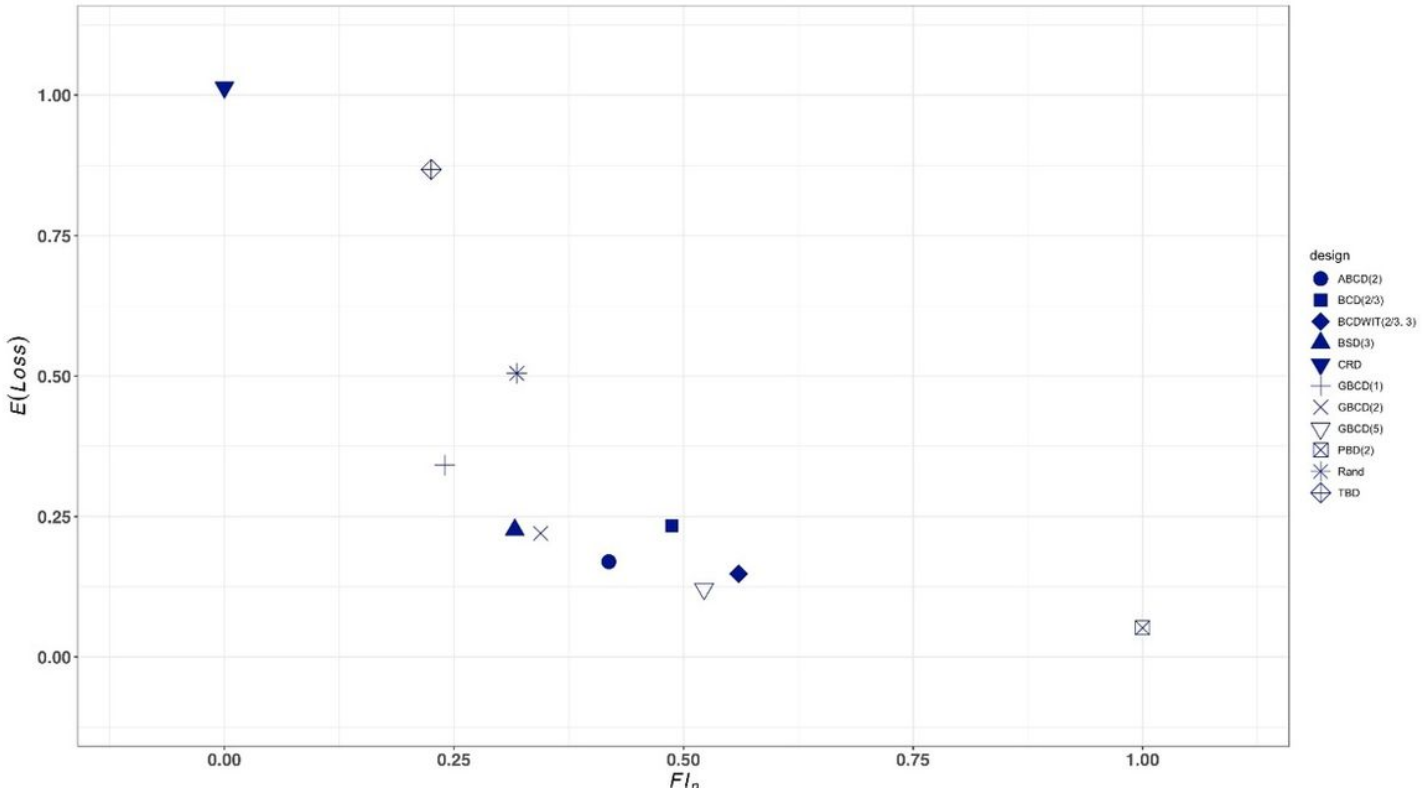


Figure 2

Simulated forcing index (x-axis) vs. aggregate expected loss (y-axis) for 11 restricted randomization procedures for n=50.

Heat map plot of the balance/randomness tradeoff metric $d(m) = \sqrt{\{Imb(m)\}^2 + \{FI(m)\}^2}$ vs. allocation step ($m = 1, \ldots, 50$) for 11 restricted randomization procedures.

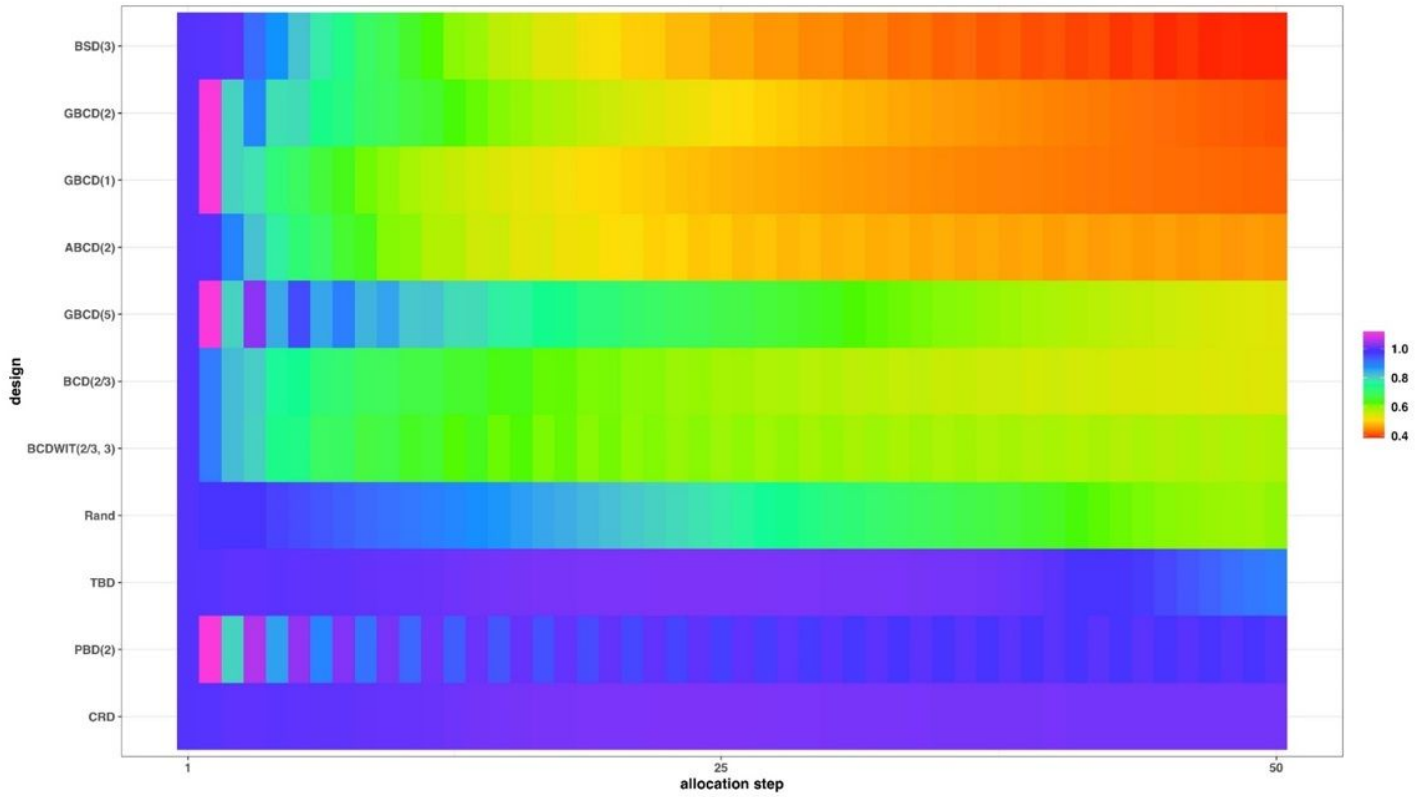


Figure 3

Heat map plot

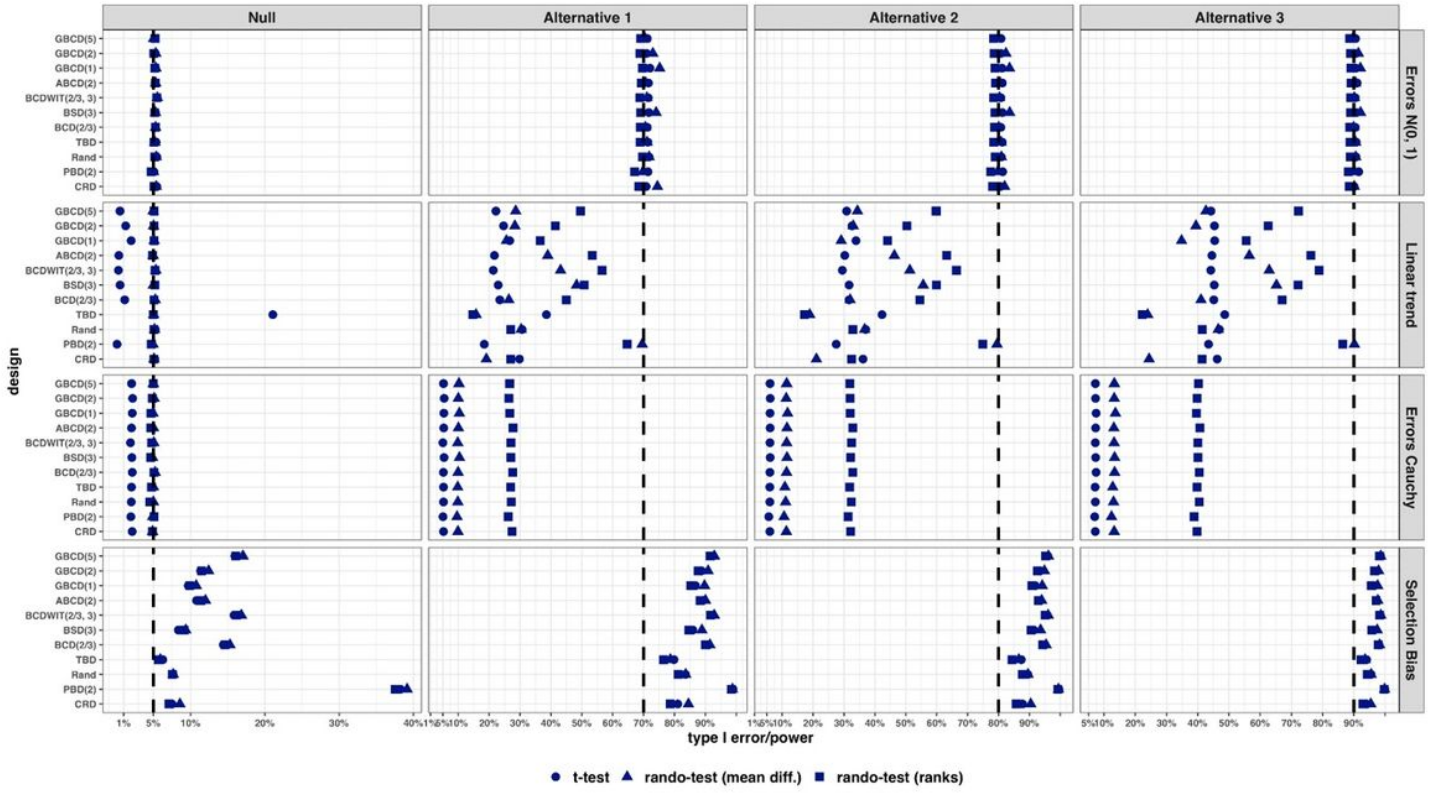


Figure 4

Simulated type I error rate and power of 11 restricted randomization procedures. Four models for the data generating mechanism of the primary outcome (M1: Normal random sampling; M2: Linear trend; M3: Errors Cauchy; and M4: Selection bias). Four scenarios for the treatment mean difference (Null; Alternatives 1, 2, and 3). Three statistical tests (T1: two-sample t-test; T2: randomization-based test using mean difference; T3: randomization-based test using ranks)

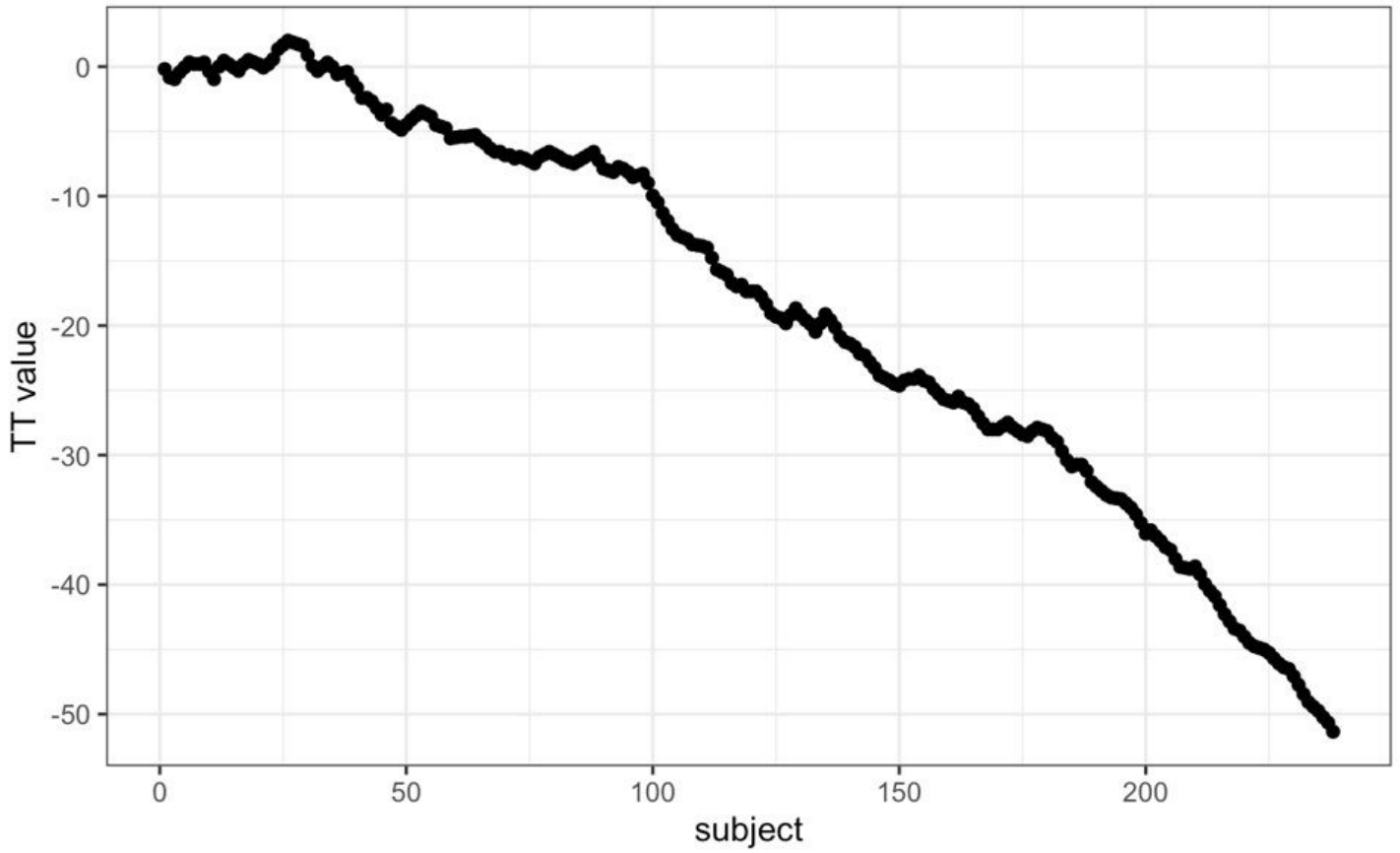


Figure 5

Cusum plot of baseline log serum bilirubin level of 248 subjects from the azathioprine trial, reproduced from Figure 1 of Altman and Royston (1988)