Pair-switching rerandomization

Ke Zhu | Hanzhong Liu

Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China

Correspondence
Hanzhong Liu, Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China. Email: lhz2016@tsinghua.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 12071242; Guo Qiang Institute of Tsinghua University

Abstract
Rerandomization discards assignments with covariates unbalanced in the treatment and control groups to improve estimation and inference efficiency. However, the acceptance-rejection sampling method used in rerandomization is computationally inefficient. As a result, it is time-consuming for rerandomization to draw numerous independent assignments, which are necessary for performing Fisher randomization tests and constructing randomization-based confidence intervals. To address this problem, we propose a pair-switching rerandomization (PSRR) method to draw balanced assignments efficiently. We obtain the unbiasedness and variance reduction of the difference-in-means estimator and show that the Fisher randomization tests are valid under PSRR. Moreover, we propose an exact approach to invert Fisher randomization tests to confidence intervals, which is faster than the existing methods. In addition, our method is applicable to both nonsequentially and sequentially randomized experiments. We conduct comprehensive simulation studies to compare the finite-sample performance of the proposed method with that of classical rerandomization. Simulation results indicate that PSRR leads to comparable power of Fisher randomization tests and is 3–23 times faster than classical rerandomization. Finally, we apply the PSRR method to analyze two clinical trial datasets, both of which demonstrate the advantages of our method.

KEYWORDS
causal inference, clinical trials, experimental design, Metropolis–Hastings algorithm, randomization-based inference, sequential experiment

1 | INTRODUCTION

Randomized experiments are the gold standard for drawing causal inference because all observed and unobserved confounders are balanced on average by randomization. However, even in completely randomized experiments, there is a high chance of producing an unbalanced assignment (Rosenberger and Sverdlov, 2008; Rubin, 2008; Xu and Kalbfeisch, 2010), and the probability of producing such an unbalanced assignment increases with the number of covariates (Krieger et al., 2019; Morgan and Rubin, 2012). Some researchers advocate balancing covariates in the design stage to improve the credibility of a study and increase the precision of estimation and power of tests (e.g., Greevy et al., 2004; Harshaw et al., 2020; Kallus, 2018; Student, 1938). Fisher (1926) proposed to use stratification to balance a few categorical covariates that are most relevant to the outcomes. When there are many important covariates, a large class of covariate-adaptive or covariate-adjusted response-adaptive randomization methods has been proposed to achieve covariate balance across treatment groups by sequentially modifying the probabilities of assignments (Hu and Hu, 2012; Hu et al., 2014; Pocock and Simon, 1975; Rosenberger and Sverdlov, 2008; Taves, 1974). If the covariates of all units can be collected before the physical implementation of experiments, such as in
many phase 1 clinical trials (Senn, 2013) and randomized controlled trials in economics (Athey and Imbens, 2017; Banerjee et al., 2020), units can be repeatedly assigned to treatment groups until all categorical and continuous covariates are satisfactorily balanced, which is often called rerandomization (RR). For example, Maclure et al. (2006) used RR in a clinical trial to assess the influence of physician education tools (PETs) on the prescription quality for general practitioners. According to a survey (Bruhn and McKenzie, 2009), RR is commonly used in the design stage of policy evaluation to balance key covariates but is less commonly documented (Heckman and Karapakula, 2021).

Although balanced designs can improve efficiency, scholars advocate against deterministic designs that sacrifice randomness for the sake of better balance (Efron, 1971; Johansson et al., 2021; Kapelner et al., 2021). If we force the exact balance within each (small) block in a small trial, knowing past allocations allows for the accurate prediction of most future allocations in the same block, which may introduce selection bias into the trial (Rosenberger and Lachin, 2015, Section 5). Moreover, deterministic designs do not enjoy the merits of randomization as a reasoned basis for inference (Rosenberger and Lachin, 2015, Section 6). For further critiques of deterministic designs, we refer readers to Rosenberger and Sverdlov (2008) and Senn (2013).

Unlike deterministic designs, RR maintains randomization and can be seen as “a harmony of optimal deterministic design and completely randomized design” (Kapelner et al., 2021). Morgan and Rubin (2012) formally investigated the theoretical properties of RR using the Mahalanobis distance. This RR procedure only accepts those assignments with the Mahalanobis distance of the covariate means between the treatment and control groups less than or equal to a prespecified threshold. They derived the validity and efficiency gains of RR under the conditions of additive treatment effects and equal treatment and control group sizes. Moreover, they proposed the use of Fisher randomization tests (Fisher, 1935) to take the RR into account in the analysis stage and demonstrated its validity. Under more general conditions, Li et al. (2018) derived the asymptotic distribution of the difference-in-means estimator under RR and proposed an asymptotically valid inference method without imposing any parametric modeling assumptions on the potential outcomes and covariates.

When the sample size was small or moderate, the simulation studies in Johansson et al. (2021) showed that under RR, the asymptotic test proposed by Li et al. (2018) failed to control the type one error. In contrast, the Fisher randomization tests performed well. In general, many scholars advocate Fisher randomization tests as a credible and flexible inference approach over asymptotic inference based on extensive empirical studies (Bind and Rubin, 2020; Keele, 2015; Proschans and Dodd, 2019; Young, 2019) and theoretical analyses (Branson, 2021; Cohen and Fogarty, 2022; Caughhey et al., 2021; Luo et al., 2021; Wu and Ding, 2021; Zhao and Ding, 2021).

However, it is computationally challenging to perform Fisher randomization tests under RR (Luo et al., 2021). In fact, Fisher randomization tests are computationally intensive, even under complete randomization (CR) (Bind and Rubin, 2020; Chung et al., 2018). Under RR, generating only one well-balanced assignment often requires drawing thousands of assignments, not to mention, sampling numerous well-balanced assignments to conduct Fisher randomization tests. The computational issue is even more severe when we construct randomization-based confidence intervals by inverting a series of Fisher randomization tests (Imbens and Rubin, 2015, Section 5.7).

Modern sampling techniques, such as Markov chain Monte Carlo, help improve the sampling efficiency (Givens and Hoeting, 2013; Liu, 2008). Motivated by the Metropolis–Hasting algorithm (Hastings, 1970; Metropolis et al., 1953), our first contribution is to propose a PSRR method, which can save a huge amount of computational cost of classical RR. Our main idea is to sample an acceptable (well-balanced) assignment along the path of gradual improvement in covariate balance. Specifically, we start with a completely randomized assignment. If it is not acceptable, then we try to move, or “rerandomize,” to a more balanced assignment by “pair-switching;” that is, switching the treatment status of two randomly selected units—one is in the treatment group, and the other is in the control group. To prevent trapping in a local optimum, we allow moving to less-balanced assignments with specific probabilities. We switch until we find an acceptable assignment. Our simulations and real data analysis illustrate that PSRR is approximately 22–23 times faster than classical RR to achieve comparable powers of Fisher randomization tests. Furthermore, we obtain the unbiasedness and a lower bound on the variance reduction of the difference-in-means estimator under PSRR.

Most relevant to our study, Krieger et al. (2019) proposed a greedy pair-switching (GPS) algorithm to improve the balance performance of classical RR. Rather than sampling an assignment under which covariate balance is achieved (the Mahalanobis distance is less than a prespecified threshold), GPS tries to find an assignment with locally optimal balance by GPS. In particular, in the case of dividing n units into equal-sized treatment and control groups, GPS starts with a random assignment and then moves to the most balanced assignment of all \((n/2)^2\) tentative pair-switching assignments until no pair-switching can further reduce the imbalance. In contrast, PSRR allows us to move to less-balanced assignments with specific probabilities. Thus, PSRR is a nongreedy heuristic, which shares elements with other heuristics, such as the
simulated annealing algorithm (Givens and Hoeting, 2013, Section 3.3) and epsilon-greedy algorithm (Sutton and Barto, 2018, Section 2.2). By sacrificing short-term benefits, the nongreedy heuristic explores the space more fully and prevents trapping in the local optimum. In our simulation, PSRR only tries two to five times to move to a new candidate assignment and is approximately 4–110 times faster than GPS to achieve comparable powers of Fisher randomization tests.

Our second contribution is to propose an exact approach to construct randomization-based confidence intervals by inverting Fisher randomization tests. Existing methods determine the endpoints of the intervals using numerical approximations (Garthwaite, 1996; Wang and Rosenberger, 2020; Luo et al., 2021). In contrast, our approach finds the endpoints by solving a series of linear equations, whose computational cost is almost the same as that of a single Fisher randomization test.

In many clinical trials and A/B testing applications (Bertsimas et al., 2019; Bhat et al., 2020; Kapelner and Krieger, 2014; Qin et al., 2016), the units enroll the experiment sequentially. The experimenter might be unable to wait to conduct the experiment until all the experimental units arrive. This motivates researchers to consider the sequential assignment of treatment status one by one (fully sequential experiments) or group by group (group sequential experiments). To improve the estimation and inference efficiency in group sequential experiments, Zhou et al. (2018) proposed a sequential RR (SeqRR) method. However, SeqRR is plagued by low computational efficiency. Our third contribution is the generalization of PSRR in group sequential experiments. Sequential PSRR is three to seven times faster than SeqRR with comparable powers.

The remainder of this paper is organized as follows. In Section 2, we introduce the potential outcomes framework, RR, Fisher randomization tests, and randomization-based confidence intervals. In Section 3, we propose a PSRR method and study its theoretical properties. We extend our method to sequentially randomized experiments in Section 4. In Section 5, we conduct simulation studies to compare the performance of PSRR with that of existing methods. In Section 6, we illustrate our method using two clinical trial datasets. We conclude the paper with a discussion of future work in Section 7. All proofs and additional simulation results are provided in Web Appendices.

2 RANDOMIZED EXPERIMENTS AND RERANDOMIZATION

2.1 Framework and notation

To ground our discussion, we introduce a phase 1 clinical trial assessing the interactions between oral reserpine and intravenous methamphetamine. The goal of this study was to evaluate the safety of reserpine for the withdrawal of methamphetamine dependence. After all participants were recruited, 20 were randomly assigned to the treatment group, which received reserpine plus methamphetamine, and the remaining 10 participants received placebo plus methamphetamine (control group). The outcomes of interest were a series of pharmacological responses, including posttreatment heart rate, which we use for illustration purposes. Summary statistics of the dataset are shown in Table C7 in Web Appendix C. The posttreatment heart rate of the treatment group was significantly lower than that of the control group, which is contrary to the prediction of the protocol. Moreover, some baseline covariates were not well balanced in the sense that the standardized differences were outside the range $[-0.1,0.1]$ (Austin, 2009). In particular, the pretreatment heart rate of the treatment group was lower than that of the control group. Balancing these covariates may be helpful in improving the interpretability of this study and the estimation efficiency of the treatment effect.

In this study, we adopt the Neyman–Rubin potential outcomes framework to define the treatment effect (Neyman et al., 1990; Rubin, 1974). For the $i$th participant in the trial ($i = 1, ..., n$), we suppose that she/he has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, which represent her/his posttreatment heart rates receiving reserpine plus methamphetamine and placebo plus methamphetamine, respectively. The unit-level treatment effect is defined as $\tau_i = Y_i(1) - Y_i(0)$. As each participant can only receive reserpine plus methamphetamine or placebo plus methamphetamine, but not both, we can only observe one of the potential outcomes. Thus, the unit-level treatment effect is not identifiable without other modeling assumptions. Fortunately, under the stable unit treatment value assumption (SUTVA) (Rubin, 1980), which states that the potential outcomes of one unit are unaffected by the treatment status of other units and that there is only one version of each treatment status, we can estimate the average treatment effect, which is defined as $\tau = n^{-1} \sum_{i=1}^{n} (Y_i(1) - Y_i(0))$. For each participant $i$, let us denote $W_i$ as her/his treatment status, $W_i = 1$ if he/she received reserpine plus methamphetamine and $W_i = 0$ otherwise. We denote the number of units in the treatment group as $n_t$, and the number of units in the control group as $n_c$. Let $W = (W_1, ..., W_n)^T$ be the vector of treatment assignment. In completely randomized experiments, the probability that $W$ takes a particular value $w = (w_1, ..., w_n)$ with $w_i \in \{0, 1\}$ is $P(W = w) = n! n_c!/n!$, $\sum_{i=1}^{n} w_i = n_t$. Under SUTVA, the observed outcome of unit $i$ is $Y_i = W_iY_i(1) + (1 - W_i)Y_i(0)$. For each participant $i$, we also observe a $p$-dimensional vector of baseline covariates $X_i = (X_{i1}, ..., X_{ip})^T$, such as age, sex, and pretreatment heart rate. The covariate matrix is denoted by $X = (X_{11}, ..., X_{n1})^T$. These covariates can be either continuous or categorical.
Let $\overline{X} = n^{-1} \sum_{i=1}^{n} X_i$. Our goal is to infer $\tau$ using observed data $\{Y_i, W_i, \{X_i\}_{i=1}^{n}\}$.

### 2.2 | Re-randomization

Morgan and Rubin (2012) suggested using the Mahalanobis distance of the covariate means in the treatment and control groups to measure the covariate balance. For a given assignment $W$, the Mahalanobis distance is defined as

$$M(W) \equiv \left(\overline{X}_t - \overline{X}_c\right)^T \left[\text{cov}\left(\overline{X}_t - \overline{X}_c\right)\right]^{-1} \left(\overline{X}_t - \overline{X}_c\right),$$

where $\overline{X}_t = \sum_{i: W_i = 1} X_i/n_t$ and $\overline{X}_c = \sum_{i: W_i = 0} X_i/n_c$ are the mean vectors of the covariates in the treatment and control groups, respectively, and $S_{XX} = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T$ is the covariation matrix of the covariates. The treatment assignment is acceptable if $M(W) \leq a$, where $a > 0$ is a prespecified threshold. The whole procedure of RR is as follows:

1. Collect covariates data and specify the balance criterion as $M(W) \leq a$.
2. Randomly generate a candidate assignment with $n_t$ units in the treatment group and $n_c$ units in the control group.
3. Check whether the candidate assignment is acceptable by the balance criterion. If acceptable, proceed to Step 4. Otherwise, return to Step 2.
4. Conduct the experiment using the acceptable assignment.

We denote the set of acceptable assignments by $\mathcal{W}_a = \{W : M(W) \leq a, \sum_{i=1}^{n} W_i = n\}$. When $a = \infty$, RR is equivalent to CR and $\mathcal{W}_\infty = \{W : \sum_{i=1}^{n} W_i = n\}$. The above procedure can be regarded as an acceptance-rejection sampling method: samples from $\mathcal{W}_\infty$ are accepted if they fall into $\mathcal{W}_a \subset \mathcal{W}_\infty$, and rejected otherwise. Morgan and Rubin (2012) proved that the asymptotic acceptance probability of the assignment is $p_a = P(X_p^2 < a)$. In practice, Li et al. (2018) recommended that $p_a = 0.001$, which means that we need to draw approximately $1/p_a = 1000$ treatment assignments in $\mathcal{W}_\infty$ to find an acceptable assignment in $\mathcal{W}_a$.

### 2.3 | Fisher randomization tests

As discussed in Section 1, after conducting the experiment using the assignment generated by RR, we can use Fisher randomization tests to test the sharp null hypothesis, $H_0 : Y_i(1) - Y_i(0) = 0, i = 1, ..., n$. Under $H_0$, we can impute all potential outcomes by using the observed outcomes, $\tilde{Y}_i(1) = \tilde{Y}_i(0) = Y_i$. We first select a test statistic to test $H_0$. A widely used test statistic is the difference-in-means estimator, $\hat{\tau}(W) = \sum_{i: W_i = 1} Y_i(1) - \sum_{i: W_i = 0} Y_i(0) / n_c$. Under $H_0$, the exact distribution of $\hat{\tau}(W)$ is known; thus, we can calculate the exact $p$-value. However, the computation of the exact $p$-value is intensive when the cardinality of set $\mathcal{W}_a$, denoted by $|\mathcal{W}_a|$, is too large. In practice, we often use the Monte Carlo method to approximate the exact $p$-value. More specifically, we independently sample $B$ assignments $\{W^b, b = 1, ..., B\}$ following the RR procedure and then calculate the corresponding values of the test statistic $\{\hat{\tau}(W^b), b = 1, ..., B\}$ under $H_0$. For the two-sided alternative $H_1 : Y_i(1) - Y_i(0) \neq 0$, larger values of $|\hat{\tau}(W^b)|$ indicate a departure from $H_0$ in favor of $H_1$. Thus, the approximated $p$-value is defined as the proportion of $|\hat{\tau}(W^b)|$ values larger than or equal to the observed $|\hat{\tau}(W^{obs})|$: that is, $\hat{p} = B^{-1} \sum_{b=1}^{B} I\{|\hat{\tau}(W^b)| \geq |\hat{\tau}(W^{obs})|\}$, where $I\{\cdot\}$ is an indicator function and $\hat{\tau}(W^{obs})$ is the observed assignment. To obtain a good approximation of the $p$-value, we need $B$ to be sufficiently large, for example, $B = 1000$. When $p_a = 0.001$, we need to draw, on average, $B/p_a = 10^6$ treatment assignments from $\mathcal{W}_\infty$ to obtain $B = 1000$ acceptable assignments and then perform Fisher randomization tests with an accuracy of the $p$-value of approximately $1/B = 10^{-3}$.

### 2.4 | Randomization-based confidence intervals

Because of the dual relationship between hypothesis testing and interval estimation, we can construct randomization-based confidence intervals by inverting Fisher randomization tests (Lehmman, 1963). We first consider the construction of a lower confidence bound with confidence level $1 - \alpha$, $0 < \alpha < 1$. For this purpose, we need to test $H_0^2 : Y_i(1) - Y_i(0) = \theta, i = 1, ..., n$, versus $H_1^2 : Y_i(1) - Y_i(0) > \theta, i = 1, ..., n$, where $\theta \in \mathbb{R}$ is a hypothetical average treatment effect. Under $H_0^2$, we impute the unobserved outcomes and denote the observed and imputed outcomes as $Y^{imp}_\theta = \{\tilde{Y}_i(1), \tilde{Y}_i(0), i = 1, ..., n\}$. For any hypothetical $W^b \in \mathcal{W}_a$, the corresponding value of the test statistic is denoted by $\hat{\tau}(Y^{imp}_\theta, W^b)$. Then, the exact $p$-value function is defined as $p(\theta) = |\mathcal{W}_a|^{-1} \sum_{W^b \in \mathcal{W}_a} I\{|\hat{\tau}(Y^{imp}_\theta, W^b)| \geq |\hat{\tau}(Y^{imp}_\theta, W^{obs})|\}$. Letting $\hat{\theta}_l \equiv \sup\{\theta : p(\theta) \leq \alpha\}$, Luo et al. (2021, Proposition 2) showed that $\hat{\theta}_l$ is a lower confidence bound; that is, $[\hat{\theta}_l, \infty)$ covers the true average treatment effect $\tau$ with a probability of at least $1 - \alpha$. Several approaches have been proposed to approximate $\hat{\theta}_l$, such as the grid search.
method (Imbens and Rubin, 2015; Rosenbaum, 2002), Robbins–Monro algorithm (Garthwaite, 1996; Wang and Rosenberger, 2020), and bisection method (Wang and Rosenberger, 2020; Luo et al., 2021). In this study, we propose an exact approach to obtain the lower confidence bound, which saves significant computational costs.

The existing methods only use the monotonicity of \( p(\theta) \). However, for a test statistic that is nondecreasing with respect to \( \theta \), such as the difference-in-means estimator, \( p(\theta) \) is also a right-continuous step function with finite jump points. If we can exactly determine all these jump points, we can recover \( p(\theta) \) and obtain the exact lower confidence bound \( \hat{\theta}_l \). The observed value of the test statistic \( \hat{\tau}(\hat{Y}) \) does not change with \( \theta \). Moreover, for any hypothetical assignment \( W^b \in \mathcal{W}_a \) and \( W^b \neq W^{\text{obs}} \), the value of the test statistic \( \hat{\tau}^{\text{obs}}(\theta) = \hat{\tau}(Y^{\text{imp}}_\theta, W^{\text{obs}}) \) increases with \( \theta \). As \( p(\theta) \) is the proportion of \( \hat{\tau}(\theta) \) greater than or equal to \( \hat{\tau}^{\text{obs}} \), \( p(\theta) \) jumps only if one of the \( \hat{\tau}(\theta) \)'s that is smaller than \( \hat{\tau}^{\text{obs}} \) becomes equal to \( \hat{\tau}^{\text{obs}} \). Therefore, the jump point of \( p(\theta) \) is \( \theta \) such that \( \hat{\tau}(\theta) \) is equal to \( \tau^{\text{obs}} \).

**Theorem 1.** For \( W^b \in \mathcal{W}_a \), \( W^b \neq W^{\text{obs}} \), the solution of \( \hat{\tau}(Y^{\text{imp}}_\theta, W^b) = \hat{\tau}(Y^{\text{imp}}_\theta, W^{\text{obs}}) \) is:

\[
\hat{\theta}_b = \frac{\sum_{W^{\text{obs}}_i=1} W^b_i=0 Y_i(1) - \sum_{W^{\text{obs}}_i=0} W^b_i=1 Y_i(0)}{\sum_{i=1}^n I_{W^{\text{obs}}_i=1,W^b_i=0}}.
\]

For \( W^b = W^{\text{obs}} \), let \( \hat{\theta}_b = -\infty \). Let \((\hat{\theta}_1, \ldots, \hat{\theta}(|\mathcal{W}_a|))\) be the increasingly ordered value of \( \hat{\theta}_b \). Subsequently, the one-sided \( 1 - \alpha \) lower confidence bound \( \hat{\theta}_l \) is \( \hat{\theta}_l([\alpha]|\mathcal{W}_a|+1) \).

When \(|\mathcal{W}_a|\) is too large, we have to use the Monte Carlo approximation of the exact \( p \)-value function. Specifically, we independently sample \( B \) assignments from \( \mathcal{W}_a \) and approximate the exact \( p \)-value function by \( \hat{p}(\theta) = B^{-1} \sum_{W^b \in \mathcal{W}^{MC}_a} I_{\{\hat{\tau}(Y^{\text{imp}}_\theta, W^b) \geq \hat{\tau}(Y^{\text{imp}}_\theta, W^{\text{obs}})\}} \), where \( \mathcal{W}^{MC}_a \) is the set of assignments generated by the Monte Carlo approximation. Then, we can approximate \( \hat{\theta}_l \) using \( \hat{\theta}_l \equiv \sup[\theta : \hat{p}(\theta) \leq \alpha] \), which is obtained by replacing \( \mathcal{W}_a \) in Theorem 1 with \( \mathcal{W}^{MC}_a \).

Similarly, we can construct the one-sided \( 1 - \alpha \) upper confidence bound. The endpoints of the \( 1 - \alpha \) two-sided confidence interval are the one-sided \( 1 - \alpha \) lower and upper confidence bounds. The exact approach is not only useful for PSRR, but also applicable to other experimental designs.

### 3 Pair-switching rerandomization

Our main goal is to sample treatment assignment \( W \) from \( \mathcal{W}_a \) more efficiently. To this end, we start with a random assignment and move toward an acceptable assignment through a random walk chain, \((W^{(0)}, \ldots, W^{(T)})\). Specifically, when \( t = 0 \), we sample a completely randomized assignment \( W^{(0)} \) and compute the corresponding Mahalanobis distance \( M^{(0)} \). If \( M^{(t)} > a \), we randomly switch between one treated unit and one control unit in \( W^{(t)} \) to obtain \( W^* \) and compute \( M^* = M(W^*) \). If \( M^* \leq M^{(t)} \), we move to \( W^* \), that is, making \( t = t + 1 \) and \( W^{(t)} = W^* \). Otherwise, we move to \( W^* \) with positive probability \( (M^{(t)}/M^*)^\gamma \), which prevents us from being trapped in a local optimum. We continue to move until \( M^{(t)} \leq a \). We summarize the whole procedure of PSRR using the Mahalanobis distance in Algorithm 1. Tuning parameter \( \gamma \) controls the probabilities of movement. A larger value...
of γ results in a smaller probability of moving to a less-balanced assignment. In contrast, a smaller value of γ results in more random movements. In the extreme case of γ = ∞, we never move to a less-balanced assignment, and we randomly move regardless of the balance if γ = 0. In our simulation studies and real data analysis, the performance of PSRR is robust for a wide range of values of γ. The default value is set to γ = 10.

To further improve computational efficiency, we can use the following tricks to compute the Mahalanobis distance: First, as X is not affected by treatment assignments, we compute $S_{XX}$ only once during the entire procedure of the Fisher randomization tests. This trick is applicable to RR, GPS, and PSRR. Second, suppose we switch the i-th and j-th elements of $W^{(i)}$ with $W^{(i)} = 1$ and $W^{(j)} = 0$ and generate a new assignment $W^*$ with $W^{*(i)} = 0$ and $W^{*(j)} = 1$. If we have computed the Mahalanobis distance $M^{(i)}$ of the assignment $W^{(i)}$, we can simplify the calculation of the Mahalanobis distance of $W^*$ by

$$M(W^*) = M^{(i)} - \left(2 \sum_{l=1}^{n} W^{(i)}_l H_{ll} - H_{ll}\right) + \left(2 \sum_{l=1}^{n} W^{*(i)}_l H_{ll} - H_{ll}\right) + h_i - h_j,$$

where $H = XS_{XX}^{-1}X^T/\{n_i(1-n_i/n)\}$ and $h = (2n_i/n)H_{11}$ with 1 as an n-dimensional column vector of 1’s. The quantities $H$ and $h$ depend only on $X$; thus, we only need to compute them once. The second trick is only applicable to GPS and PSRR. The proof of (1) is provided in Web Appendix A.

In the design stage, we run Algorithm 1 once and output assignment $W_{\text{obs}}$ to conduct the experiment. In the analysis stage, we run Algorithm 1 B times independently and output $\{W^b, b = 1, \ldots, B\}$. We then follow the procedures in Section 2 to perform Fisher randomization tests and construct randomization-based confidence intervals. Because we use the same procedure to generate treatment assignments in the design and analysis stages, Fisher randomization tests preserve the significance level, and randomization-based confidence intervals have the desired coverage rate (Imbens and Rubin, 2015; Luo et al., 2021; Rosenberger and Lachin, 2015).

Similar to RR, PSRR still maintains the unbiasedness of the difference-in-means estimator $\hat{\tau}$ when $n_t = n_c$. Intuitively, every assignment $W$ has a symmetrical assignment $1 - W$, such that their Mahalanobis distances are equal. Because the distribution of assignments induced by PSRR (Algorithm 1) is driven by the Mahalanobis distance, symmetry implies that $P(W_l = 1) = P(W_l = 0) = 1/2$. This further implies unbiasedness; see the following Theorem 2.

**Theorem 2.** Suppose that $n_t = n_c = n/2$ and $W$ is generated by PSRR. Then, $E(\hat{\tau}) = \tau$.

If the covariates are related to the outcomes, then the variance of the difference-in-means estimator under PSRR is smaller than that under CR.

**Theorem 3.** If (i) $n_t = n_c = n/2$, (ii) $w_l = 0, 1$, $Y_i(w_i) = \beta_0 + \beta^T X_i + \tau w_i + e_i$, where $\beta_0 + \beta^T X_i$ is the linear projection of $Y_i(0)$ onto $(1, X)$ and $e_i$ is the deviation from the linear projection, and (iii) $\hat{\tau}$ and $\overline{X}_i - \overline{X}_c$ are normally distributed, then we have

$$\frac{\text{var}_{\text{CR}}(\hat{\tau}) - \text{var}_{\text{PSRR}}(\hat{\tau})}{\text{var}_{\text{CR}}(\hat{\tau})} \geq (1 - a/p)R^2,$$

where the subscripts CR and PSRR represent distributions under CR and PSRR, respectively, and $R^2 \equiv \beta^T \text{cov}_{\text{CR}}(\overline{X}_i - \overline{X}_c)\beta/\text{var}_{\text{CR}}(\hat{\tau})$.

Morgan and Rubin (2012) assumed similar conditions to obtain the variance reduction of RR, which is equal to $\left(1 - v_{a,RR}\right)R^2$ with $v_{a,RR} = P(\chi^2_{p+2} \leq a)/P(\chi^2_{p} \leq a)$. Because PSRR is more complicated than RR, we only obtain a lower bound for the variance reduction. Compared to $\left(1 - v_{a,RR}\right)R^2$, the bound $\left(1 - a/p\right)R^2$ is smaller, but very close to it. For instance, when $p = 10$, $R^2 = 0.5$, and $p_0 = 0.001$, the former is equal to 44% and the latter is equal to 42.6%. Moreover, the lower bound becomes increasingly tight as $p$ increases.

**4 | SEQUENTIAL PAIR-SWITCHING RERANDOMIZATION**

In this section, we generalize PSRR to sequentially randomized experiments. We start with a clinical trial that aimed to evaluate the efficacy of the Therapeutic Education System (TES), an Internet-delivered treatment for substance or alcohol abuse (Campbell et al., 2014). Participant recruitment information was distributed to 10 outpatient centers. Patients interested in this project were referred to the researchers for screening. If the patients were eligible, they proceeded to subsequent baseline measurements and randomization. After 15 months, 507 eligible patients were sequentially recruited and randomized immediately upon arrival or within 1 month. Finally, $n_t = 255$ patients were allocated to the treatment group (treatment as usual plus TES), and $n_t = 252$ patients were allocated to the control group (treatment as usual). The primary outcome was the
abstinence from drug or heavy alcohol use, assessed using urine drug tests and self-reports. Some baseline covariates were considered to be strong predictors of the outcome, such as the number of days from the participants last drug or alcohol use. In this trial, because the patients received usual therapy regardless of assignment to the treatment or control group, the experimenters could wait for a while (e.g., a month) to recruit a group of patients and perform randomization together.

Similar to the nonsequentially randomized experiment, Zhou et al. (2018) proposed a SeqRR method to balance the baseline covariates. In the following, we first review the basic concepts of SeqRR and then propose a sequential PSRR (SeqPSRR) procedure to reduce the computational cost.

Assume that \( n \) patients are divided into \( K \) sequential groups of sizes \( n_1, ..., n_K \). In group \( k \) ( \( k = 1, ..., K \) ), patients are randomly assigned to the treatment group and the remaining \( n_{k} = (1 - e) n_{k} \) patients are assigned to the control group, where \( e \in (0,1) \) is the propensity score. We denote the covariates in the first \( k \) groups as \( X_{1:k} \), whose dimension is \( (n_1 + \cdots + n_k) \times p \). We denote the covariance matrix of \( X_{1:k} \) as \( S_{XX}[k] \). There are two main differences between sequential and classical RRs: (1) when we assign treatment to patients in the former groups, we cannot access covariate data in the latter groups; (2) when we assign treatment to patients in the latter groups, we cannot change the treatment assignment of the patients in the former groups, although we still have to consider covariate balance of all arrived groups.

SeqRR proceeds as follows: For the first group, we randomly assign \( n_{11} \) patients to the treatment group and the other \( n_{1c} \) to the control group. This assignment is denoted by vector \( W_{[1]} = (W_{1}, ..., W_{n_1}) \). The Mahalanobis distance corresponding to assignment \( W_{[1]} \) is defined as follows:

\[
M_1(W_{[1]}) \equiv (\bar{X}_{[1]} - \bar{X}_{[1]})^T \text{cov} (\bar{X}_{[1]} - \bar{X}_{[1]})^{-1} (\bar{X}_{[1]} - \bar{X}_{[1]}),
\]

where \( \bar{X}_{[1]} = \sum_{i \in \{W_i = 1\}} X_i/n_{11} \) and \( \bar{X}_{[1]} = \sum_{i \in \{W_i = 0\}} X_i/n_{1c} \) are the mean vectors of the covariates under treatment and control in the first group, respectively. If \( M_1(W_{[1]}) \leq a_1 \), a prespecified threshold, we accept the assignment \( W_{[1]} \) and conduct the experiment for the patients in the first group. Otherwise, we rerandomize until \( M_1(W_{[1]}) \leq a_1 \) for some \( W_{[1]} \). If \( K = 1 \), this step is the same as that in the classical RR.

When the patients in the \( k \)th group enroll in the experiment, we randomly assign \( n_{ik} \) patients to the treatment group and the other \( n_{ck} \) patients to the control group. This assignment is denoted by a vector \( W_{[k]} = (W_{n_1:k+1}, ..., W_{n_{k+1}}) \), where \( n_{1:k} = \sum_{i=1}^{K} n_i \) is the total number of patients in the first \( k \) groups. Then, the Mahalanobis distance corresponding to assignment \( W_{[k]} \) is defined as:

\[
M_k(W_{[k]}) \equiv (\bar{X}_{[1:k]} - \bar{X}_{[1:k]})^T \text{cov} (\bar{X}_{[1:k]} - \bar{X}_{[1:k]})^{-1} (\bar{X}_{[1:k]} - \bar{X}_{[1:k]})
\]

\[
= n_{1:k} (1 - n_{1:k}/n_k) (\bar{X}_{[1:k]} - \bar{X}_{[1:k]})^T S_{XX}[k]^{-1} (\bar{X}_{[1:k]} - \bar{X}_{[1:k]}),
\]

where \( n_{1:k} = \sum_{i=1}^{k} n_i, n_{c1:k} = \sum_{i=1}^{k} n_{ci} \) are the total numbers of patients in the treatment and control arms in the first \( k \) groups, and \( \bar{X}_{[1:k]} = \sum_{i : W_i = 1} X_i/n_{1:k} \) and \( \bar{X}_{[1:k]} = \sum_{i : W_i = 0} X_i/n_{c1:k} \) are the mean vectors of the covariates in the treatment and control arms in the first \( k \) groups, respectively. If \( M_k(W_{[k]}) \leq a_k \), we accept assignment \( W_{[k]} \) and conduct the experiment for the patients in the \( k \)th group. Otherwise, we rerandomize the \( n_k \) units in the \( k \)th group until \( M_k(W_{[k]}) \leq a_k \) for some \( W_{[k]} \).

Zhou et al. (2018) provided suggestions on the choice of \( a_k \). They derived the optimal allocation of the expected number of draws in group \( k \), denoted by \( s_k \), based on asymptotic arguments (see Proposition 2 therein). Moreover, they showed that the conditional distribution of \( M_k | M_{k-1} \) is a noncentral chi-square. We can then calculate \( a_k \) based on \( s_k \) and \( M_{k-1} \).

Zhou et al. (2018) also showed the unbiasedness of the difference-in-means estimator \( \hat{\tau} \) for estimating the average treatment effect \( \tau \) given \( n_{ik} = n_{ck}, k = 1, ..., K \). Although a proper asymptotic inference procedure has not been established under SeqRR, Fisher randomization tests are still valid. Like classical RR, the computational cost of Fisher randomization tests under SeqRR is also very high, motivating us to consider sequential PSRR (SeqPSRR).

The SeqPSRR procedure is presented in Algorithm 2. We sequentially replace classical RR with PSRR for each group. When \( K = 1 \), Algorithm 2 reduces to Algorithm 1. We also use the tricks discussed in Section 3 to compute the Mahalanobis distance. We run Algorithm 2 once and output \( W_{obs} = (W_{[1]}, ..., W_{[K]}) \) sequentially to conduct the experiment. We sample \( B \) assignments \( \{W^b, b = 1, ..., B\} \) following Algorithm 2 independently, perform Fisher randomization tests, and construct randomization-based confidence intervals following the procedures introduced in Section 2.
ALGORITHM 2 Sequential pair-switching rerandomization

\begin{algorithm}
\KwIn{Expected numbers of draws in each group, \(s_1, \ldots, s_K\), and tuning parameter \(\gamma\) (default value \(\gamma = 10\)).}
Set \(M[0] = 0\);
Set \(n_0 = 0\);
\For{\(k = 1, \ldots, K\)}{
\KwIn{Covariates data \(X_{[1:k]}\).}
Set \(a_k = n_k(n_1) - 1 q_k\), where \(q_k\) is the lower \(1/s_k\) quantile of a non-central chi-square distribution with \(p\) degrees of freedom and a non-central parameter \(n_{1:(k-1)}(n_k)^{-1} M_{[k-1]}\); 
Set \(t = 0\);
Set \(W_{[k]}^{(0)}\) as \(n_{tk}\) elements equal to 1 and \(n_{ck}\) elements equal to 0 with random positions;
Set \(M_{[k]}^{(0)} = M_k(W_{[k]}^{(0)})\);
\While{\(M_{[k]}^{(t)} > a_k\)}{
Randomly switch the positions of one of the 1’s and one of the 0’s in \(W_{[k]}^{(t)}\) and obtain \(W_{[k]}^{*}\);
Set \(M_{[k]}^{*} = M_k(W_{[k]}^{*})\);
Sample \(J\) from a Bernoulli distribution with probability \(\min\{\left(M_{[k]}^{(t)}/M_{[k]}^{*}\right)^{\gamma}, 1\}\);
\If{\(J = 1\)}{
Set \(t = t + 1\);
Set \(W_{[k]}^{(t)} = W_{[k]}^{*}\);
Set \(M_{[k]}^{(t)} = M_{[k]}^{*}\);
}\end\While
Set \(M_{[k]} = M_{[k]}^{(t)}\);
\KwOut{\(W_{[k]} = W_{[k]}^{(t)}\).}
\end\algorithm

By applying the arguments in Theorem 2 to each group separately, we show that the SeqPSRR also maintains the unbiasedness of the difference-in-means estimator when \(n_{tk} = n_{ck}\) for all \(k = 1, \ldots, K\).

Theorem 4. Suppose that \(n_{tk} = n_{ck} = n_{k}/2, k = 1, \ldots, K\), and \(W\) is generated from SeqPSRR. Then, \(E(\bar{\tau}) = \tau\).

Zhou et al. (2018) showed that, compared to CR, the proportion of variance reduction of SeqRR is \(\{1 - E(M_K) / p\}R^2\). Under the more complicated SeqPSRR, we can obtain a lower bound for the proportion of variance reduction. Following Morgan and Rubin (2012) and Zhou et al. (2018), we assume equal treatment group sizes, additive treatment effects, and normally distributed difference-in-means of the covariates and outcomes.

Theorem 5. If (i) \(n_{tk} = n_{ck} = n_{k}/2, k = 1, \ldots, K\), (ii) for \(w_i = 0, 1\), \(Y_i(w_i) = \beta_0 + \beta'X_i + \tau w_i + e_i\), where \(\beta_0 + \beta'X_i\) is the linear projection of \(Y_i(0)\) onto \((1, X)\) and \(e_i\) is the deviation from the linear projection, and (iii) \(\bar{\tau}\) and \(\bar{X}_i - \bar{X}_c\) are normally distributed, then we have

\[
\frac{\text{var}_{CR}(\bar{\tau}) - \text{var}_{SeqPSRR}(\bar{\tau})}{\text{var}_{CR}(\bar{\tau})} \geq (1 - a_K / p)R^2,
\]

where the subscript SeqPSRR represents the distribution under sequential PSRR.
5 | SIMULATION STUDIES

5.1 | Nonsequentially randomized experiments

We compare PSRR with existing methods in both nonsequentially and sequentially randomized experiments. To fairly compare the speed of these methods, we use R to implement all methods. For each case, we run the code on an Intel Xeon E5-2690 V4 processor (2.6GHz, 35M Cache, 28 Core, 128G Memory). We replicate the simulation $n_{rep} = 1000$ times to examine the repeated sampling properties.

In this section, we consider nonsequentially randomized experiments. The covariates are generated from the standard normal distribution, $X_{ij} \stackrel{i.i.d.}{\sim} N(0,1)$, where i.i.d. stands for “independent and identically distributed.” Similar to the simulation setups in Johansson et al. (2021), the potential outcomes $Y_i(0), i = 1, \ldots, n$, are generated independently by a linear regression model, $Y_i(0) = X_{i1} + \cdots + X_{ip} + \epsilon_i$, where $\epsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$. We set $n = 30, 50, 100, p = 10$, and choose $\sigma^2$ such that $R^2 = \text{var}(X_{i1} + \cdots + X_{ip})/\text{var}(Y_i(0)) = 0.2$ or 0.5. The $R^2$ measures the correlations between covariates and potential outcomes. Usually, the larger the $R$-square, the greater the benefit of balancing covariates (Morgan and Rubin, 2012).

To examine the size (type one error) and power (one minus type two error) of the Fisher randomization tests, we, respectively, set $Y_i(1) = Y_i(0)$ and $Y_i(1) = Y_i(0) + 0.3\sqrt{\text{var}(Y_i(0))}$. Both covariates and potential outcomes are generated once and then kept fixed. The results for larger sample sizes, $n = 500, 1000, 2000$, are similar and provided in Web Appendix C.

We consider four design methods: CR, GPS (Krieger et al. (2019)), RR (Morgan and Rubin (2012)), and PSRR. We consider equal-sized treatment and control groups. For the two RR methods, the threshold $\alpha$ of the Mahalanobis distance satisfies $p_\alpha = 0.001$, following the recommendations of Li et al. (2018). For PSRR, we set the default value of the tuning parameter $\gamma$ to 10. We also examine the performance of PSRR with different values of $\gamma$. The results are similar and provided in Web Appendix C. We use three randomness metrics to measure the randomness of these methods, see Web Appendix B. We perform Fisher randomization tests with a significance level $\alpha = 0.05$ and construct randomization-based confidence intervals with a nominal coverage rate of 95%. We set $B = 1000$. We use both the proposed exact approach and bisection method (Luo et al., 2021; Wang and Rosenberger, 2020) to invert the Fisher randomization tests to confidence intervals and compare their performances.

Table 1 shows the results of the nonsequentially randomized experiments when $R^2 = 0.5$. The results for $R^2 = 0.2$ are provided in Web Appendix C. First, the difference-in-means estimator $\hat{\tau}$ under all design methods has negligible finite-sample biases. Second, compared with CR, the other three design methods reduce the standard deviation of $\hat{\tau}$ by 25%–31%, 22%–25%, and 22%–27%, respectively. Notably, compared with RR and PSRR, although GPS improves the balance of assignments, it does not significantly reduce the variance of $\hat{\tau}$. This is because when the Mahalanobis distance is already small, the gain in efficiency by further pursuing balance is often negligible (Johansson et al., 2021). Third, under all design methods and sample sizes, Fisher randomization tests control the type one error, and randomization-based confidence intervals reach the nominal coverage rate. Fourth, compared to CR, the other three design methods increase the power by 57%–91%, 60%–68%, and 38%–63%, respectively, and reduce the interval length by 22%–28%, 21%–24%, and 20%–24%, respectively. Because the bisection method produces almost the same (but slightly wider) interval as the exact approach, we do not present its coverage probability and interval length owing to space restrictions. Fifth, PSRR dramatically reduces the computational costs: it is 4–110 times faster than GPS and 22–23 times faster than RR. Moreover, with an increase in $n$, the computation time of GPS increases dramatically. In contrast, the computation times of RR and PSRR decrease as $n$ increases because the number of iterations used to find each acceptable assignment decreases, as shown in Figure 1. The improvement in the speed of PSRR relative to RR is not strongly dependent on the sample size. Finally, compared to the computational time required to generate the desired assignments, the time required to construct confidence intervals using the proposed exact approach is negligible (approximately 0.02 s), whereas the computational time of the bisection method is nonnegligible (approximately 0.6 s) under both CR and PSRR.

Figure 1 shows that PSRR requires much smaller total iterations than GPS and RR to find an acceptable assignment, which is why PSRR is much faster than the other two methods. Because the total number of iterations is equal to the average number of inner iterations (the number of attempts before jumping to a new candidate assignment) times the number of outer iterations (the number of candidate assignments before finding an acceptable assignment), we further compare the inner and outer iterations of these three methods. Because GPS enumerates all pair-switched assignments of the current assignment and jumps to the best, its inner iterations are equal to $(n/2)^2$, which increases rapidly as $n$ increases. It takes approximately five to seven outer iterations to find an acceptable assignment (i.e., to reach a local optimum). RR has no inner loop and requires approximately 1296–2435 outer iterations to find one acceptable assignment. PSRR
takes approximately two to five attempts to jump to a new candidate assignment, which is much smaller than that of GPS, and 15–20 outer iterations to find one acceptable assignment, which is much smaller than that of RR. Overall, PSRR outperforms both GPS and RR in terms of the number of total iterations (39–70 compared to 1087–17,245 and 1296–2435, respectively), which leads to PSRR sampling acceptable assignments more quickly.

5.2 | Sequentially randomized experiments

In this section, we consider sequentially randomized experiments. The experimental units are recruited in $K = 3, 5, 10$ groups, and each group has $n_k = 20$ units. We also consider the setting of $n_k = 50$ and provide the results in Web Appendix C. We simulate the covariates and potential outcomes in the same way as those in Section 5.1. We consider three design methods: sequential CR (SeqCR), SeqRR, and sequential PSRR (SeqPSRR). For the two RR methods, we set $S = 1000$ and determine $s_k$ and $a_k$ following the strategy proposed by Zhou et al. (2018). Specifically, we set $(s_1, s_2, s_3) = (30, 136, 834)$, when $K = 3$; $(s_1, \ldots, s_5) = (10, 10, 29, 133, 818)$ when $K = 5$; and $(s_1, \ldots, s_{10}) = (10, 10, 10, 10, 10, 10, 10, 10, 10, 28, 128, 774)$, when $K = 10$. Because $a_k$ is determined based on chi-square approximation, to find an assignment such that $M_k \leq a_k$, the actual number of RRs is probably much greater than $s_k$ in the finite sample. We follow Zhou et al. (2018) to allow at most $10s_k$ RRs and output the best assignment if all these RRs are not acceptable. We
further examine the influence of the 10s_\k constraint in Web Appendix C.

Table 1 presents the performance of all methods when \( R^2 = 0.5 \). The results for \( R^2 = 0.2 \) are provided in Web Appendix C. The overall conclusions are similar to those in the nonsequentially randomized experiments. Under all three designs, the biases of \( \hat{\tau} \) are negligible, Fisher randomization tests control for type one error, and randomization-based confidence intervals have the desired coverage rate. Compared with SeqCR, the other two methods, SeqRR and SeqPSRR, reduce the standard deviation of the difference-in-means estimator by 25%–33%, increase the power by 50%–82%, and reduce the interval length by 26%–30%. SeqPSRR is three to seven times faster than SeqRR. The improvement in speed decreases as \( K \) increases, because as \( K \) becomes larger, the \( s_k \)'s become smaller in most groups, and smaller \( s_k \) values directly reduce the improvement in speed (in the extreme case with \( s_k = 1 \), SeqCR, SeqRR, and SeqPSRR are equivalent in group \( k \)). Overall, SeqPSRR achieves comparable precision of point and interval estimation and power of tests, with less computational cost than SeqRR.

6 | CLINICAL TRIAL EXAMPLES

6.1 | A phase 1 clinical trial

We revisit the nonsequentially randomized experiment introduced in Section 2 and illustrate the applicability and advantages of PSRR. In this experiment, \( n_t = 20 \) and \( n_c = 10 \). To balance the eight important covariates using RR and PSRR, we set the threshold \( a \) as the 0.001 quantile of \( \chi^2_8 \), which leads to \( a = 0.86 \). In contrast, the Mahalanobis distance corresponding to the actual assignment is 10.95, the 0.795 quantile of \( \chi^2_8 \), which is a rather unbalanced assignment.

We generate 10,000 assignments by CR, RR, and PSRR. RR takes approximately 19 min, whereas PSRR takes approximately 1 min. Figure 2(a) shows the empirical distributions of the standardized differences in covariate means and the empirical percent reductions in variance (PRIVs) relative to CR. PSRR performs similarly to RR in terms of balancing covariates. The empirical PRIVs for the covariates are close to the theoretical lower bound \( 100(1 - a/p)\% \approx 89.3\% \). According to Theorem 3, the lower bound
FIGURE 2  Box-plot of standardized differences in covariate means for two datasets. The dashed lines indicate the recommended univariate balance thresholds [−0.1, 0.1] (Austin, 2009). The diamonds indicate the standardized differences in covariate means for the actual assignment in the experiment. PRIV stands for the empirical percent reductions in variance. For each covariate, the methods from top to bottom are CR, RR, and PSRR for subplot (A) and SeqCR, SeqRR, and SeqPSRR for subplot (B), respectively. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

of the PRIV for the treatment effect estimation is \( \{100(1 - \frac{a}{p})R^2\}\% \approx 38.6\% \), where \( R^2 = 0.432 \) is the adjusted R-square obtained by regressing the outcome on the covariates. To examine this conclusion, a simulation based on semi-synthetic data is provided in Web Appendix C. If we use the assignment generated by PSRR to replace the original assignment in the experiment, all covariates will be well balanced in the design stage. Thus, the difference in posttreatment heart rate will no longer be attributable to the baseline difference.

6.2 An Internet-delivered clinical trial

We consider the sequentially randomized experiment introduced in Section 4 and show the applicability and
advantages of SeqPSRR. We split \( n = 507 \) participants into \( K = 15 \) sequential groups (with an approximate waiting time of 1 month per group) with sizes \( n_1 = 31 \) and \( n_k = 34, \ 2 \leq k \leq 15 \). We set \( n_{15} = 17, \ 1 \leq k \leq 15, \) such that \( n = 255 \). To balance the covariates of age, the number of days since the last use of drugs or alcohol, and their quadratic terms and interaction \(( p = 5)\) using SeqRR and SeqPSRR, we set \( s_k = 10 \) for \( 1 \leq k \leq 12, s_{13} = 12, s_{14} = 68, \) and \( s_{15} = 800 \).

We generate 10,000 assignments by SeqCR, SeqRR, and SeqPSRR. SeqPSRR is more than twice as fast as SeqRR (334 s versus 867 s). Figure 2(b) shows the empirical distributions of the standardized differences in the covariates and PRIVs relative to SeqCR. For the original assignment, the standardized differences in age and their quadratic terms are larger than 0.1, indicating that they are not well balanced (Austin, 2009). In contrast, both SeqRR and SeqPSRR produce balanced assignments. The empirical PRIVs for the covariates are close to the theoretical lower bound \( 100(1 - a_k/p)^2 \% \approx 99.6\% \). According to Theorem 5, the lower bound of the PRIV for the treatment effect estimation is \( 100(1 - a_k/p)R^2 \% \approx 13.8\% \), where \( R^2 = 0.139 \) is the adjusted \( R \)-square obtained by regressing the outcome on the covariates. To examine this conclusion, a simulation based on semisynthetic data is provided in Web Appendix C. Overall, SeqPSRR balances the covariates as well as SeqRR and reduces the computational cost of conducting Fisher randomization tests.

7 | DISCUSSION

RR can improve the efficiency of statistical inference by balancing baseline covariates in the design stage. However, the low sampling efficiency of classical RR forces researchers to make a trade-off between feasibility and covariate balance, which leads to inferior statistical performance. In this article, we propose PSRR and SeqPSRR to balance the baseline covariates. Compared with classical RR and SeqRR, the proposed methods can achieve comparable precision of point and interval estimates and power of tests, but with a much lower computational cost. We derive the unbiasedness and a lower bound for the variance reduction of the difference-in-means estimator under both PSRR and SeqPSRR. In addition, we propose an exact approach to invert Fisher randomization tests to construct randomization-based confidence intervals. Extensive simulation studies and two clinical trial data analyses demonstrate the advantages of the proposed methods. Under PSRR, the assignments are no longer uniformly distributed on \( W_{a_t} \). It is challenging to derive the (asymptotic) distribution of commonly used test statistics, such as the difference-in-means estimator. We leave this problem to future work. Another limitation of the proposed methods is that they are not applicable to fully sequential experiments, in which units enroll the experiments one by one.

In the clinical trial introduced in Section 2, baseline measurements were performed on days 0 and 1, randomization was performed on day 3, and actual treatment allocation was conducted on day 4. During this period, some baseline covariates, such as blood pressure, may change. However, Fisher randomization tests remain valid as long as the same covariates are used in the design and analysis stages. In short, we emphasize that “one should analyze as one designs” (Rosenberger and Lachin, 2015, Section 6.4). Since earlier values of covariates are often as predictive of outcomes as later values of covariates, balancing earlier measured covariates can also improve statistical efficiency. If we take the values that are the closest to the treatment allocation as the true values of the covariates, we can view earlier values of the covariates as measurements with errors. Wang and Ma (2021) studied the impact of measurement error on covariate-adaptive randomization. It would be interesting to extend their theory to PSRR.

In the clinical trial introduced in Section 4, 49 out of 255 participants in the TES group were reported to have not completed the entire 12-week TES course, whereas 252 participants in the control group did not have access to TES during the study. For this one-sided noncompliance (treatment switching) issue, our method provides a valid intention-to-treat (ITT) analysis of the effect of assignment (Imbens and Rubin, 2015, Chapter 23). To infer the effect of the actual receipt of treatment, Mattei et al. (2020) addressed the problem of treatment switching using principal stratification (Frangakis and Rubin, 2002) and proposed a Bayesian approach. Rubin (1998) established a framework for utilizing Fisher randomization tests in the presence of imperfect compliance. We can adopt Rubin’s framework to handle noncompliance problems in PSRR.

RR balances covariates in the design stage. Another approach to dealing with covariate imbalance is to use regression adjustment in the analysis stage (Bloniarz et al., 2016; Lin, 2013; Liu and Yang, 2020; Lei and Ding, 2021; Su and Ding, 2021), which has also been combined with RR to further improve efficiency (Li and Ding, 2020). It would be interesting to combine the PSRR and regression-adjusted Fisher randomization tests (Zhao and Ding, 2021).

RR has been extended to experiments with multiple arms (Branson et al., 2016; Li et al., 2020), where we need to balance multiple contrasts of covariate means simultaneously. Thus, conducting Fisher randomization tests might face a more severe computational burden than RR in two-arm experiments. Therefore, it would be interesting to generalize PSRR to experiments with multiple arms, including the factorial experiments.
We use the Mahalanobis distance as a balance measure. It is straightforward to extend our methods to RR using other balance measures, such as the Mahalanobis distance within tiers of covariate importance (Morgan and Rubin, 2015), rank-based balance measure with estimated weights of the covariates (Johansson and Schultzberg, 2020), ridge RR (Branson and Shao, 2021), and PCA RR (Zhang et al., 2021).

Recently, the choice of threshold \( \alpha \) in RR has been further investigated. Kapelner et al. (2022) proposed a procedure to determine the optimal RR threshold based on a trade-off between the observed imbalance and the risk of unobserved imbalance. Banerjee et al. (2020) provided guidelines for choosing RR thresholds based on a trade-off between covariate balance and robustness. In practice, classical RR using these recommended thresholds may not be computationally feasible. PSRR addresses this issue.

**ACKNOWLEDGMENTS**

The authors are grateful to the Associate Editor and two referees for their valuable comments. Dr. Ke Zhu would like to thank Professors Donald Rubin and Ke Deng for their helpful comments during the causal inference seminar at Tsinghua University. Dr. Hanzhong Liu’s research is supported by the National Natural Science Foundation of China (Grant No. 12071242) and Guo Qiang Institute of Tsinghua University. The information reported here results from secondary analyses of data from clinical trials conducted by the National Institute on Drug Abuse (NIDA). Specifically, data from NIDA-CPU-0006 (A Phase 1 Parallel-Group, Double-Blind, Placebo-Controlled Cardiovascular and Behavioral Study Assessing Interactions Between Single Doses of Oral Reserpine and Intravenous Methamphetamine) and NIDA-CTN-0044 (Web-delivery of Evidence-Based, Psychosocial Treatment for Substance Use Disorders) were included. NIDA databases and information are available at http://datashare.nida.nih.gov.

**DATA AVAILABILITY STATEMENT**

The data that support the findings in this paper are openly available in National Institute on Drug Abuse (NIDA) databases at https://datashare.nida.nih.gov. Specifically, data from Jones (2017) and Nunes (2014) were included.

**ORCID**

Ke Zhu https://orcid.org/0000-0001-8778-0086
Hanzhong Liu https://orcid.org/0000-0002-0028-5136

**REFERENCES**

Athey, S. & Imbens, G.W. (2017) The econometrics of randomized experiments. In: Banerjee, A.V. & Duflo, E. (Eds.) Handbook of economic field experiments, volume 1. Amsterdam: Elsevier, pp. 73–140.

Austin, P.C. (2009) Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25), 3083–3107.

Banerjee, A.V., Chassang, S., Montero, S. & Snowberg, E. (2020) A theory of experimenters: robustness, randomization, and balance. *American Economic Review*, 110(4), 1206–1230.

Bertsimas, D., Kollolko, N. & Weinstein, A.M. (2019) Covariate-adaptive optimization in online clinical trials. *Operations Research*, 67(4), 1150–1161.

Bhat, N., Farias, V.F., Maallemi, C.C. & Sinha, D. (2020) Near-optimal AB Testing. *Management Science*, 66(10), 4477–4495.

Bind, M.A.C. & Rubin, D.B. (2020) When possible, report a Fisher-exact \( P \) value and display its underlying null randomization distribution. *Proceedings of the National Academy of Sciences*, 117(32), 19151–19158.

Bloniarz, A., Liu, H., Zhang, C.H., Sekhon, J.S. & Yu, B. (2016) Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27), 7383–7390.

Branson, Z. (2021) Randomization tests to assess covariate balance when designing and analyzing matched datasets. *Observational Studies*, 7(2), 1–36.

Branson, Z., Dasgupta, T. & Rubin, D.B. (2016) Improving covariate balance in \( 2^k \) factorial designs via rerandomization with an application to a New York City Department of Education High School Study. *The Annals of Applied Statistics*, 10(4), 1958–1976.

Branson, Z. & Shao, S. (2021) Ridge rerandomization: an experimental design strategy in the presence of covariate collinearity. *Journal of Statistical Planning and Inference*, 211, 287–314.

Bruhn, M. & McKenzie, D. (2009) In pursuit of balance: randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4), 200–232.

Campbell, A.N., Nunes, E.V., Matthews, A.G., Stitzer, M., Miele, G.M., Polsky, D. et al. (2014) Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. *American Journal of Psychiatry*, 171(6), 683–690.

Caughey, D., Dafoe, A., Li, X. & Miratrix, L. (2021) Randomization inference beyond the sharp null: bounded null hypotheses and quantiles of individual treatment effects. *arXiv preprint arXiv:2101.09195*.

Chung, M.K., Wang, Y., Huang, S.G. & Lyu, I. (2018) Rapid acceleration of the permutation test via slow random walks in the permutation group. *arXiv preprint arXiv:1812.06696*.

Cohen, P.L. & Fogarty, C.B. (2022) Gaussian prepivoting for finite population causal inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2), 295–320.

Efron, B. (1971) Forcing a sequential experiment to be balanced. *Biometrika*, 58(3), 403–417.

Fisher, R. (1926) The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33, 503–515.

Fisher, R.A. (1935) *The design of experiments* (1st ed.). Edinburgh: Oliver and Boyd.

Frangakis, C.E. & Rubin, D.B. (2002) Principal stratification in causal inference. *Biometrics*, 58(1), 21–29.
Garthwaite, P.H. (1996) Confidence intervals from randomization tests. Biometrics, 52(4), 1387–1393.
Givens, G.H. & Hoeting, J.A. (2013) Computational Statistics. Hoboken, NJ: John Wiley & Sons.
Greevy, R., Lu, B., Silber, J.H. & Rosenbaum, P. (2004) Optimal multivariate matching before randomization. Biostatistics, 5(2), 263–275.
Harshaw, C., Sävje, F., Spielman, D. & Zhang, P. (2020) Balancing covariates in randomized experiments using the Gram-Schmidt Walk. arXiv preprint arXiv:1911.03071.
Hastings, W. (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1), 97–109.
Heckman, J.J. & Karapakula, G. (2021) Using a satisficing model of experimenter decision-making to guide finite-sample inference for compromised experiments. The Econometrics Journal, 24(2), C1–C39.
Hu, F., Hu, Y., Ma, Z. & Rosenberger, W.F. (2014) Adaptive randomization for balancing over covariates. Wiley Interdisciplinary Reviews: Computational Statistics, 6(4), 288–303.
Hu, Y. & Hu, F. (2012) Asymptotic properties of covariate-adaptive randomization. The Annals of Statistics, 40(3), 1794–1815.
Imbens, G.W. & Rubin, D.B. (2015) Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge: Cambridge University Press.
Johansson, P., Rubin, D.B. & Schulzberg, M. (2021) On optimal rerandomization designs. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2), 395–403.
Johansson, P. & Schulzberg, M. (2020) Rerandomization strategies for balancing covariates using pre-experimental longitudinal data. Journal of Computational and Graphical Statistics, 29(4), 1–26.
Jones, R. (2017) A phase 1 parallel-group, double-blind, placebo-controlled cardiovascular and behavioral study assessing inter-actions between single doses of oral reserpine and intravenous methamphetamine. National Institute on Drug Abuse (NIDA) Data Share web site.
Kallus, N. (2018) Optimal a priori balance in the design of controlled experiments. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(1), 85–112.
Kapelner, A. & Krieger, A. (2014) Matching on-the-fly: sequential allocation with higher power and efficiency. Biometrics, 70(2), 378–388.
Kapelner, A., Krieger, A.M., Sklar, M. & Azriel, D. (2022) Optimal rerandomization designs via a criterion that provides insurance against failed experiments. Journal of Statistical Planning and Inference, 219, 63–84.
Kapelner, A., Krieger, A.M., Sklar, M., Shalit, U. & Azriel, D. (2021) Harmonizing optimized designs with classic randomization in experiments. The American Statistician, 75(2), 195–206.
Keele, L. (2015) The statistics of causal inference: a view from political methodology. Political Analysis, 23(3), 313–335.
Krieger, A.M., Azriel, D. & Kapelner, A. (2019) Nearly random designs with greatly improved balance. Biometrika, 106(3), 695–701.
Lehmann, E.L. (1963) Nonparametric confidence intervals for a shift parameter. The Annals of Mathematical Statistics, 34(4), 1507–1512.
Lei, L. & Ding, P. (2021) Regression adjustment in completely randomized experiments with a diverging number of covariates. Biometrika, 108(4), 815–828.
Li, X. & Ding, P. (2020) Rerandomization and regression adjustment. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(1), 241–268.
Li, X., Ding, P. & Rubin, D.B. (2018) Asymptotic theory of rerandomization in treatment–control experiments. Proceedings of the National Academy of Sciences, 115(37), 9157–9162.
Li, X., Ding, P. & Rubin, D.B. (2020) Rerandomization in 2k factorial experiments. The Annals of Statistics, 48(1), 43–63.
Lin, W. (2013) Agnostic notes on regression adjustments to experimental data: Reexamining fremdman’s critique. The Annals of Applied Statistics, 7(1), 295–318.
Liu, H. & Yang, Y. (2020) Regression-adjusted average treatment effect estimates in stratified randomized experiments. Biometrika, 107(4), 935–948.
Liu, J.S. (2008) Monte Carlo strategies in scientific computing. Springer Science & Business Media.
Luo, X., Dasgupta, T., Xie, M. and Liu, R.Y. (2021) Leveraging the Fisher randomization test using confidence distributions: inference, combination and fusion learning. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(4), 777–797.
Maclure, M., Nguyen, A., Carney, G., Dormuth, C., Roelants, H., Ho, K. et al. (2006) Measuring prescribing improvements in pragmatic trials of educational tools for general practitioners. Basic & Clinical Pharmacology & Toxicology, 98(3), 243–252.
Mattei, A., Mealli, F. & Ding, P. (2020) Assessing causal effects in the presence of treatment switching through principal stratification. arXiv preprint arXiv:2002.11989.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equation of state calculations by fast computing machines. The Journal of Chemical Physics, 21(6), 1087–1092.
Morgan, K.L. & Rubin, D.B. (2012) Rerandomization to improve covariate balance in experiments. The Annals of Statistics, 40(2), 1263–1282.
Morgan, K.L. & Rubin, D.B. (2015) Rerandomization to balance tiers of covariates. Journal of the American Statistical Association, 110(512), 1412–1421.
Neyman, J., Dabrowska, D.M. & Speed, T.P. (1990) On the application of probability theory to agricultural experiments. Statistical Science, 5(4), 465–472.
Nunes, E.V. (2014) Web-delivery of evidence-based, psychosocial treatment for substance use disorders. National Institute on Drug Abuse (NIDA) Data Share web site.
Pocock, S.J. & Simon, R. (1975) Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics, 31(1), 103–115.
Proschan, M.A. & Dodd, L.E. (2019) Re-rerandomization tests in clinical trials. Statistics in Medicine, 38(12), 2292–2302.
Qin, Y., Li, Y., Ma, W. & Hu, F. (2016) Pairwise sequential randomization and its properties. arXiv preprint arXiv:1611.02802.
Rosenbaum, P.R. (2002) Observational Studies (2nd ed.). New York: Springer.
Rosenberger, W.F. & Lachin, J.M. (2015) Randomization in clinical trials: theory and practice. Hoboken, NJ: John Wiley & Sons.
Rosenberger, W.F. & Sverdlov, O. (2008) Handling covariates in the design of clinical trials. Statistical Science, 23(3), 404–419.
Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688–701.
Rubin, D.B. (1980) Randomization analysis of experimental data: the Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371), 591–593.

Rubin, D.B. (1998) More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine*, 17(3), 371–385.

Rubin, D.B. (2008) Comment: the design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484), 1350–1353.

Senn, S. (2013) Seven myths of randomisation in clinical trials. *Statistics in Medicine*, 32(9), 1439–1450.

Student, (1938) Comparison between balanced and random arrangements of field plots. *Biometrika*, 29(3–4), 363–378.

Su, F. & Ding, P. (2021) Model-assisted analyses of cluster-randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5), 994–1015.

Sutton, R.S. & Barto, A.G. (2018) *Reinforcement learning: an introduction*. Massachusetts: MIT Press.

Taves, D.R. (1974) Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics*, 15(5), 443–453.

Wang, T. & Ma, W. (2021) The impact of misclassification on covariate-adaptive randomized clinical trials. *Biometrics*, 77(2), 451–464.

Wang, Y. & Rosenberger, W.F. (2020) Randomization-based interval estimation in randomized clinical trials. *Statistics in Medicine*, 39(21), 2843–2854.

Wu, J. & Ding, P. (2021) Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, 116(536), 1898–1913.

Xu, Z. & Kalbfleisch, J.D. (2010) Propensity score matching in randomized clinical trials. *Biometrics*, 66(3), 813–823.

Young, A. (2019) Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2), 557–598.

Zhang, H., Yin, G. & Rubin, D.B. (2021) PCA rerandomization. *arXiv preprint arXiv:2102.12262*.

Zhao, A. & Ding, P. (2021) Covariate-adjusted Fisher randomization tests for the average treatment effect. *Journal of Econometrics*, 225(2), 278–294.

Zhou, Q., Ernst, P.A., Morgan, K.L., Rubin, D.B. & Zhang, A. (2018) Sequential rerandomization. *Biometrika*, 105(3), 745–752.

**SUPPORTING INFORMATION**

Web Appendices referenced in Sections 1–6 and R codes to reproduce the numerical results are available with this paper at the Biometrics website on Wiley Online Library.

**How to cite this article:** Zhu, K., & Liu, H. (2023) Pair-switching rerandomization. *Biometrics*, 79, 2127–2142. [https://doi.org/10.1111/biom.13712](https://doi.org/10.1111/biom.13712)