Received: 8 April 2019 | Accepted: 11 January 2021

DOI: 10.1111/rssb.12417

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

On optimal rerandomization designs

Per Johansson[1,2] | **Donald B. Rubin**[2] | **Mårten Schultzberg**[3]

[1]Uppsala University and IFAU, Uppsala, Sweden

[2]Tsinghua University, Beijing, China

[3]Uppsala University, Uppsala, Sweden

Correspondence
Per Johansson, Uppsala University and IFAU, Uppsala, Sweden.
Email: per.johansson@statistics.uu.se

Abstract

Blocking is commonly used in randomized experiments to increase efficiency of estimation. A generalization of blocking removes allocations with imbalance in covariate distributions between treated and control units, and then randomizes within the remaining set of allocations with balance. This idea of *rerandomization* was formalized by Morgan and Rubin (*Annals of Statistics*, 2012, **40**, 1263–1282), who suggested using Mahalanobis distance between treated and control covariate means as the criterion for removing unbalanced allocations. Kallus (*Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 2018, **80**, 85–112) proposed reducing the set of balanced allocations to the minimum. Here we discuss the implication of such an 'optimal' rerandomization design for inferences to the units in the sample and to the population from which the units in the sample were randomly drawn. We argue that, in general, it is a bad idea to seek the optimal design for an inference because that inference typically only reflects uncertainty from the random sampling of units, which is usually hypothetical, and not the randomization of units to treatment versus control.

KEYWORDS

causal inference, exact inference, Mahalanobis distance, optimal design, rerandomization

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society

1 | INTRODUCTION

Randomized controlled trials (RCTs) are traditionally seen as the gold standard for causal inference. Even though the estimators from well-conducted experiments are unbiased in expectation, the estimates from any single experiment may still be far from any 'truth' due to an unlucky, albeit random, allocation. Traditionally, blocking has been used to eliminate some bad allocations. Blocking (i.e. stratification) partitions experimental units into groups (strata) based on their similarity on covariates, and then randomization is performed within each stratum, thereby ensuring that units within each stratum will be 'fairly' represented in both the treatment and control groups; see for example Imbens and Rubin (2015) for a recent overview. An alternative, or complement, to blocking that has received attention lately is to utilize modern computational capabilities by using rerandomization (see e.g. Kallus, 2018; Kapelner et al., 2018; Krieger et al., 2019; Lauretto et al., 2017; Morgan & Rubin, 2012).

More specifically, the idea of rerandomization is to remove from consideration allocations with imbalance in observed covariates between treated and control units and then randomize within the set of allocations with balance on these covariates; call the set of all allocations $\mathcal{A}$ and the set of acceptable allocations $\mathcal{A}_a$. Such a procedure is a rerandomization design with a prespecified imbalance criterion, $\mathcal{A} \rightarrow \mathcal{A}_a$, where different criteria give rise to different rerandomization designs. The idea of rerandomization, which, although previously mentioned in the literature, apparently was first formalized by Morgan and Rubin (2012), who used the Mahalanobis distance as the criterion for defining $\mathcal{A}_a$ with the aim to make inference about the sample average treatment effect (SATE), typically estimated by the difference in sample means between treated and control units.

Kallus (2018) suggested finding the allocation that minimizes the estimated sampling variance of the SATE estimator; call the set of allocations that achieve this minimum $\mathcal{A}_{Opt}$. The primary focus of this paper is to discuss the implication of using such 'optimal' rerandomization designs for inferences to the population average treatment effect (PATE) and the SATE. We focus on Mahalanobis-based rerandomization to clarify the original rerandomization idea in Morgan and Rubin (2012), and because it has well-established properties discussed below.

Section 2 reviews the background for Mahalanobis-based rerandomization. Section 3 discusses inference to the sample and to the population, as well as how to conduct inference given an optimal design. The paper concludes with a discussion in Section 4.

2 | BASIC RESULT FROM MORGAN AND RUBIN (2012) ON MAHALANOBIS-BASED RERANDOMIZATION

Following the notation in Morgan and Rubin (2012), consider a RCT with n units in the sample, indexed by i, with n_1 to be assigned to treatment and n_0 to be assigned to control; for simplicity, $n_1 = n_0$. Let $W_i = 1$ or $W_i = 0$ if unit i is assigned treatment or control, respectively, and define $\mathbf{W} = (W_1, \ldots, W_n)'$. Furthermore, let $\mathbf{X}$ be the $n \times K$ matrix of fixed covariates in the sample $(\mathbf{x}_i, i = 1, \ldots, n)$, with observed sample covariance $cov(\mathbf{X})$.

Because $n_1 = n_0$, there are $\binom{n}{n_1} = A$ possible treatment allocation (assignment) vectors labelled $\mathbf{W}^j = (W_1^j, \ldots, W_n^j)', j = 1, \ldots, A$, where $\text{card}(\mathcal{A}) = A$, that is, the cardinality of the set $\mathcal{A}$ of possible allocations. The Mahalanobis distance for allocation j is

$$M(\mathbf{W}^j, \mathbf{X}) = M^j = \frac{4}{n}\left(\hat{\boldsymbol{\tau}}_X^{j\prime} cov(\mathbf{X})^{-1} \hat{\boldsymbol{\tau}}_X^j\right), j = 1, \ldots, A,$$

where

$$\hat{\boldsymbol{\tau}}_X^j = \frac{1}{n_1}\sum_{i=1}^{n} W_i^j \mathbf{x}_i' - \frac{1}{n_0}\sum_{i=1}^{n}(1-W_i^j)\mathbf{x}_i' = \overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j.$$

Morgan and Rubin (2012) proposed accepting the *j*th allocation when its treatment assignment vector $\mathbf{W}^j$ satisfies

$$M^j \leq a,$$

where a is a positive constant.

By the central limit theorem, and supported by experience with real examples for moderate n, the sample means of the covariates will be approximately normally distributed across random samples, so that $M^j \sim \chi_K^2$ (Morgan & Rubin, 2012). Letting

$$p_a = \Pr(\chi_K^2 \leq a) \simeq \Pr(M^j \leq a), \tag{1}$$

we see that a is determined from the choice of p_a. Because the number of rerandomizations is geometrically distributed, the expected number of randomizations needed to obtain an acceptable allocation is $1/p_a$, which means that, for instance, with $p_a = 0.001$, the average number of randomizations before drawing an allocation that fulfils the criterion is 1000.

Morgan and Rubin (2012) show that when $M^j \sim \chi_K^2$, then, due to the spherical symmetry of the multivariate normal distribution,

$$Cov(\overline{\mathbf{X}}_T^j - \overline{\mathbf{X}}_C^j \,|\, \mathbf{X}, M(\mathbf{W}^j, \mathbf{X}) < a) = \nu_a Cov(\overline{\mathbf{X}}_T - \overline{\mathbf{X}}_C \,|\, \mathbf{X}), \tag{2}$$

with

$$\nu_a = \frac{\Pr(\chi^2_{(K+2)} \leq a)}{\Pr(\chi_K^2 \leq a)}; 0 < \nu_a < 1. \tag{3}$$

This result implies that the variance of the covariate mean differences across allocations in $\mathcal{A}_a$ is reduced relative to its variance across the allocations in $\mathcal{A}$ by the factor ν_a, and the percent reduction in variance of each of the covariates in $\mathbf{X}$ (or of any linear combination of them) is equal to $100(1-\nu_a)$.

Let $Y_i(w)$ be the potential outcome under treatment w for unit i. Under the stable unit treatment value assumption (SUTVA, Rubin, 1980), the observed outcome when i is assigned W_i is equal to $Y_i = (1-W_i)Y_i(0) + W_i Y(1)$. The mean difference estimator is defined as

$$\hat{\tau} = \overline{Y}_1 - \overline{Y}_0 \tag{4}$$

where $\overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n} W_i Y_i(1)$ and $\overline{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n}(1-W_i)Y_i(0)$.

Let $\hat{\tau}^{CR}$ and $\hat{\tau}^{RR}$ be the estimators defined in Equation (4) under complete randomization (i.e. when the W_i are randomly drawn from $\mathcal{A}$) and Mahalanobis-based rerandomization (i.e. when the W_i are randomly drawn from $\mathcal{A}_a$), respectively. These estimators are unbiased for SATE and also for PATE under random sampling of the n units from the population. For convenience, we also define the corresponding estimators for a specific sample s: $\hat{\tau}_s^{CR}$ and $\hat{\tau}_s^{RR}$, respectively.

Let $\mathbf{Y}(w) = (Y_1(w), Y_2(w), \ldots, Y_n(w))'$, $w = 0, 1$, and let R^2 be the squared multiple correlation between $\mathbf{Y}(0)$ and $\mathbf{X}$. Under the assumption that (i) the residual in the linear projection of $\mathbf{Y}(0)$

on $\mathbf{X}$ is normally distributed and that (ii) treatment effects are additive (so that R^2 is also the squared multiple correlation between $\mathbf{Y}(1)$ and $\mathbf{X}$), the percentage reduction in variance (PRIV) of $\hat{\tau}_s^{RR}$ and $\hat{\tau}^{RR}$ versus the corresponding estimators under complete randomization is

$$PRIV = \frac{V_n(\hat{\tau}_s^{CR}) - V_n(\hat{\tau}_s^{RR})}{V_n(\hat{\tau}_s^{CR})} = \frac{V_n(\hat{\tau}^{CR}) - V_n(\hat{\tau}^{RR})}{V_n(\hat{\tau}^{CR})} = 100 \times R^2(1 - \nu_a), \tag{5}$$

where $V_n(.)$ denotes the variance of the estimators (Morgan & Rubin, 2012). From this expression together with Equations (1) and (3), it becomes clear that the variance reduction from Mahalanobis-based rerandomization relative to complete randomization is decreasing in p_a, the strictness of the rerandomization criterion, and non-increasing in K, the dimension of $\mathbf{X}$. For additional properties of Mahalanobis-based rerandomization, see Morgan and Rubin (2012) and Li et al. (2018).

3 | THE SPECIAL CASE OF MINIMIZING THE VARIANCE

The mean difference *estimate* for sample s and allocation j is

$$\hat{\tau}_s^j = \overline{Y}_{1s}^j - \overline{Y}_{0s}^j, \tag{6}$$

where $\overline{Y}_{1s}^j = \frac{1}{n_1} \sum_{i=1:i\in s}^{n_1} W_i^j Y_i(1)$ and $\overline{Y}_{0s} = \frac{1}{n_0} \sum_{i=1:i\in s}^{n_0} (1 - W_i^j) Y_i(0)$. The SATE for sample s is

$$SATE_s = \frac{1}{n} \sum_{i\in s} (Y_i(1) - Y_i(0)) = \frac{1}{A} \sum_{j\in\mathcal{A}} \hat{\tau}_s^j = \frac{1}{A_a} \sum_{j\in\mathcal{A}_a} \hat{\tau}_s^j,$$

where $A_a = \text{card}(\mathcal{A}_a)$, and thus estimators $\hat{\tau}_s^{CR}$ and $\hat{\tau}_s^{RR}$ are unbiased for $SATE_s$ (Imbens & Rubin, 2015; Morgan & Rubin, 2012). For sample s, the variance of the estimators, $V_n(\hat{\tau}_s^{CR})$ and $V_n(\hat{\tau}_s^{RR})$, only depend on the treatment assignment mechanisms that is the experimental designs.

For the super population, as assumed in Kallus (2018),

$$PATE = E[Y(1) - Y(0)],$$

where the expectation is over all random samples with fixed sample sizes. The variance of the estimators for inference to PATE are

$$V_n(\hat{\tau}^{CR}) = EV_n(\hat{\tau}_s^{CR}) + Var(SATE). \tag{7}$$

and

$$V_n(\hat{\tau}^{RR}) = EV_n(\hat{\tau}_s^{RR}) + Var(SATE),$$

respectively. The first term of the variance decomposition (7) is the expected variance of the estimator, and the second term is simply the variance of SATE across random samples. Clearly, only the first term differs between the two designs. From these results it follows that, in line with Kallus (2018), an optimal rerandomization design for the inferences to PATE should minimize the first term in Equation (7). Optimal designs are obtained by minimizing the maximum conditional variance of the mean difference estimator under the assumption that conditional means of the outcomes under treatment and control can be estimated (Kallus, 2018). Under this assumption, the pure strategy optimal design (PSOD) finds a

single optimal allocation, that is, the allocation in the second stage is deterministic unless there are many allocations achieving the minimum.

Inspired by Cochran and Cox (1957), section 4.36 and Kallus (2018) constructed example 1, to illustrate theorem 1 in Wu (1981), or theorem 1 in Kallus (2018), which are the basis for his optimal design. These theorems show that complete randomization is minimax if the covariates $\mathbf{X}$ are independent of the outcome. In this example, $\mathbf{X}$ and Y are deterministically generated according to

$$\begin{aligned}\mathbf{X} &= (2^0, 2^1, 2^2, \ldots, 2^{n/2-1}, -(2^0), -(2^1), -(2^2), \ldots, -(2^{n/2-1}))\\ Y_i(0) &= (-1)^{X_i} - \tau/2\\ Y_i(1) &= Y_i(0) + \tau\end{aligned}$$

where $i = 1, \ldots, n$.† Note that with this data generating process, $\mathbf{X}$ has a different distribution for each n, that is larger n not only give a larger sample but the larger sample arises from a distribution with larger variance.

Let $\widehat{\tau}^B$ and $\widehat{\tau}^P$ be the mean difference estimators under blocking and pairwise-matched allocations (see Kallus, 2018, p. 97, for details), respectively. Under this data generating process, Kallus (2018) shows that $V_n(\widehat{\tau}^{CR}) = 4/(n-1)$, $V_n(\widehat{\tau}^B) = 4/(n-8)$ and $V_n(\widehat{\tau}^B) = 4/(n-8)$ and $V_n(\widehat{\tau}^P) = 8/n$, results that are in agreement with Cochran and Cox (1957), Equation (4.3).

Using Mahalanobis-based rerandomization, the optimal design is obtained by minimizing the variance in the observed covariates $\mathbf{X}$ (Equation 2), which is obtained by letting $a \equiv \min_{\mathcal{A}} M(\mathbf{W}^j, \mathbf{X})$, which, in large samples, implies $\nu_a \simeq 0$. With an additive treatment effect, this criterion implies that the PRIV of $\widehat{\tau}^{RR}$ is equal to $100 \times R^2$.

Including only $\mathbf{X}$ (i.e. not any transformation) in the Mahalanobis distance criterion Kallus (2018) implies that $V_n(\widehat{\tau}^{RR} | p_a = 0) \equiv 4$, where $p_a = 0$ is the minimal limiting acceptance criterion—that is the variance when restricting the allocations to those with $M^j = 0$ only. The statement that $V_n(\widehat{\tau}^{RR} | p_a = 0) \equiv 4$ for all n is not correct as will be shown below. This mistake, however, exposes a useful basis for discussing the foundation for inference to the experiment's sample under randomization inference and the implication this has for an optimal design for the inference to the population.

3.1 | Inference to sample in the experiment

The mistake of Kallus (2018) stems from the incorrect assumption that the allocation $\mathbf{W} = (0, 1, 0, 1, \ldots, 0, 1)'$ *uniquely* minimizes the Mahalanobis distance for all n, and therefore the variance calculations are, incorrectly, based solely on this allocation. In fact, an experiment with $n_1 = n_0$, using the Mahalanobis distance balance measure, $\text{card}(\mathcal{A}_{Opt}) \geq 2$, that is *there exist at least two allocations, not a unique smallest Mahalanobis distance*. This follows because, using the Mahalanobis distance, every allocation has a mirror allocation with 1's and 0's exchanged but with the same imbalance. Thus, the minimum number of allocations with the smallest imbalance is two (a pair of mirror allocations). For any balance measure that fulfils the mirror property (Johansson & Schultzberg, 2019; Kapelner et al., 2019), there is no single unique best allocation.

As is shown in the online Appendix, with $n = 2$ the only pair of allocations in this example have $M^1 = M^2 = 4$, and therefore $V_2(\widehat{\tau}^{RR} | p_a = 0)$ is not defined for $n = 2$. With $n = 4$ there exists one pair of allocations with $M^j = 0$, for which $V_4(\widehat{\tau}^{RR} | p_a = 0) = 4$. For $n > 4$ there is more than one pair of

†The notation on the data generation is ambiguous in Kallus (2018). The data generating process presented here is chosen to reproduce the results in Kallus (2018).

allocations that have Mahalanobis distances equal to zero, and the variance decreases in n. With $n = 8$ and $n = 16$, $V_8(\hat{\tau}^{RR} | p_a = 0) = 4/3$ and $V_{16}(\hat{\tau}^{RR} | p_a = 0) = 4/7$, respectively. The convergence rate is, thus, slower than for $V_n(\hat{\tau}^{CR})$ but it is in contrast to the claim $V_n(\hat{\tau}^{RR} | p_a = 0) \equiv 4$ for all n.

In this special case, the only source of randomness is the randomization mechanism for assigning treatments. Under the sharp null hypothesis, for example the treatment effect is zero for all units, the value of any test-statistic is known for all possible random allocations. The exact p-value associated with any test statistic is a simple function of the percentile of the observed test-statistic in the histogram of the statistics' value across all possible allocations. This implies that theoretical asymptotic variances are helpful tools only for comparing efficiency in designs where treatment assignment is randomized within a set of allocations sufficiently large for such an asymptotic argument to be appropriate. In this case, for inference with level $\alpha = 0.05$, comparisons of variances is only useful with $n \geq 8$. With $n = 8$, it follows that $\text{card}(\mathcal{A}) = \binom{8}{4} = 70$, which implies that the smallest possible p-value in a two-sided test is less than 0.05. A variance comparison between, for example rerandomization and complete randomization should be performed for a rerandomization design where a is chosen such that $2/\text{card}(\mathcal{A}_a) \leq \alpha$, where α is the desired level of the inference. This is the reason why Morgan and Rubin (2012) states that one should not use too small a value of p_a.

Kallus (2018, p. 94) refers to the rerandomization design proposed by Morgan and Rubin (2012) as the 'historically haphazard practice of rerandomization'. To the best of our understanding, the argument for this statement seems to be based on the belief that the Mahalanobis-based rerandomization minimizes the linear projection of $\mathbf{Y}$ on $\mathbf{X}$ due to a structural assumption of a linear relation between $\mathbf{Y}$ on $\mathbf{X}$. However, as pointed out in Morgan and Rubin (2012), by including interactions and non-linear functions of covariates in the Mahalanobis distance, non-linear dependencies also can be considered in Mahalanobis-based rerandomization. To exemplify the potential importance of including transformations, we introduce both X and X^2 into the Mahalanobis distance criterion in the example above.

With both X and X^2 in the Mahalanobis distance criterion, there are no allocations with $M^j = 0$ for $n \leq 16$, and the optimal criterion is instead $p^* \equiv \min p_a : \text{card}(\mathcal{A}_{Opt}) > 0$. With $p_a = p^*$ as the criterion, card $(A_{Opt}) = 2$ for $n \leq 16$, and for these single pairs of allocations, $V_2(\hat{\tau}^{RR} | p_a = p^*) = 4$, $V_4(\hat{\tau}^{RR} | p_a = p^*) = V_8(\hat{\tau}^{RR} | p_a = p^*) = V_{16}(\hat{\tau}^{RR} | p_a = p^*) = 0$. Thus, these 'optimal' designs have smaller variance than all the other designs. However, restricting randomization to one single pair, implies that randomization inference has essentially no power. To enable inference, the rerandomization criterion must be increased, thereby allowing for 'non-optimal' allocations. For example, when the inclusion criterion is set to $p_a = 0.1$, we obtain the number of allowed allocations equal 2,8, and 1,288 for $n = 4$, 8, and 16 respectively. For these values of n,, we get $V_4(\hat{\tau}^{RR} | p_a = 0.1) = 0$,$V_8(\hat{\tau}^{RR} | p_a = 0.1) = 0.5$, and $V_{16}(\hat{\tau}^{RR} | p_a = 0.1) = 0.216$. Because $V_8(\hat{\tau}^{CR}) = 0.571$ and $V_{16}(\hat{\tau}^{CR}) = 0.267$, the PRIV for $n = 8$ and 16 is 12% and 19%, respectively. For $n = 16$, there is a sufficient number of allocations (1,288) for meaningful inferences to SATE and PATE with $p_a = 0.1$.

3.2 | Inference to the population

Li et al. (2018) show that the asymptotic distribution of the mean difference estimator after Mahalanobis-based rerandomization is generally non-normal, but rather, the asymptotic distribution is a linear combination of a normal distributed variable and a truncated normal variable. Furthermore, the asymptotic sampling variances and quantile ranges of the mean difference estimator are reduced relative to when the estimation is based on complete randomization. The result in Li et al. (2018) has the important implication that using standard asymptotic inference results with a rerandomization

design leads to sometimes highly conservative inference, with possible gains in power when using appropriate asymptotic results.

To illustrate these points in a finite sample setting, a small simulation was conducted (see the online Appendix for details). Data are generated as

$$Y_i(0) = x_{i1} + x_{i2} + x_{i3} + \epsilon_i \tag{8}$$

where x_{ij}, $j = 1, 2, 3$ and $i = 1, \ldots, n$ are independent, identically and either normally distributed with mean 0 and variance 1 (i.e. $x_{ij} \sim N(0, 1), \forall j, i$) or exponentially distributed with rate 1 (i.e. $x_{ij} \sim \exp(1), \forall j, i$). The sampling error, ϵ_i, $i = 1, \ldots, n$, is iid and $\mathcal{N}(0, \sigma^2)$, where σ^2 is chosen to obtain $R^2 = 0.2$ and 0.5 in the data generating process (8). The sample size is varied; $n = 50$, 100, 200 and 400.

Based on the mean difference estimator, the power and size (nominal level 5%) of the Fisher exact test (FRT) and the asymptotic test derived in Li et al. (2018) (denoted the LDR below) under Mahalanobis-based rerandomization (with $p_a = 0.01$) are compared to the t -test, both under complete randomization and under rerandomization. Here $\tau = 0.3\frac{\sqrt{50}}{\sqrt{n}}\sqrt{\frac{Var(\mathbf{Y}(0))}{n}}$ under the alternative.

The results are similar across distributions of the covariates. As expected, the t-test under rerandomization is highly conservative whereas the LDR test has too large size for all $n < 400$. Using the FRT and the LDR (when $n = 400$), we can see that with $R^2 = 0.5$, there is roughly 50% increase in power for both type of covariates. With $R^2 = 0.2$ there is 20% improvement in power.

Using both simulated and real data, Kallus (2018) compares the empirical variances of the mean difference estimator under different designs. Given that the asymptotic sampling distribution of the estimator is only known under the Mahalanobis criterion, it is not obvious that these comparisons of empirical variances are valid procedures for evaluating the relative efficiency of the different designs. Furthermore, for the Mahalanobis distance metric, Kallus (2018) only allows the raw covariates in the Mahalanobis criterion. It is likely that by including interactions and non-linear terms of the covariates, the variances under the Mahalanobis-based rerandomization would have been reduced as was the case in the previous example.

Schultzberg and Johansson (2020) shows that when the experimental units are randomly sampled from a super population, it is possible to draw inference to the units in this super population when choosing the best pair of allocations despite no possibility of drawing meaningful inference to the units in the sample (as illustrated in the previous example). Moreover, if the Mahalanobis criterion is used to find the best allocations, the asymptotic sampling distribution is known. However, in experiments on people, they usually choose whether or not to participate or they are selectively chosen, which means that when valid inference is the goal, it is a bad idea to choose the best pair of allocations because the inference only reflects uncertainty from random sampling. In other words, the ability to introduce a fully known stochastic mechanism in the design, from which exact inference can be based, should not be sacrificed for the usually small, often negligible, gain in efficiency achieved by choosing the best pair of allocations, rather than choosing randomly from a small set of the nearly best allocations.

4 | DISCUSSION

We discuss the implication of an 'optimal' rerandomization design for inferences to the SATE or the PATE. Rerandomization removes, from consideration, allocations with imbalance in observed covariates between treated and control units, and then randomizes within the set of allocations with

acceptable balance on these covariates. Morgan and Rubin (2012) were first to formalize a procedure for rerandomization by suggesting the Mahalanobis distance as the criterion. Kallus (2018) suggests finding the 'optimal' design for the inferences to the PATE.

The Kallus (2018) optimal designs are obtained by minimizing the maximum conditional variance of the mean difference estimator under the assumption that conditional means of the outcomes under treatment and control can be estimated. Under this assumption the PSOD aims to find a single optimal allocation from which the mean difference estimate is obtained.

For inference to SATE, the cardinality of $\mathcal{A}_a$, should be large enough to allow the exact Fisher randomization test (FRT) to have non-trivial power. A variance comparison between, for example rerandomization and complete randomization, is only meaningful when $2/\text{card}(\mathcal{A}_a) \leq \alpha$, where α is the desired level of the inference.

With random sampling in the experiment asymptotic inferences to the PATE is in theory possible when $\text{card}(\mathcal{A}_a) = 1$. However, the only criterion for which the sampling distribution of the mean difference estimator is known is the Mahalanobis criterion. Also, it is in general a bad idea to use rerandomization designs with a minimum rerandomization criterion and/or select the final assignment deterministically, as suggested in Kallus (2018), as such designs only reflect uncertainty from potential random sampling. Instead, designs 'optimal' for the inference to SATE should also be used for inferences to the units of the population. In other words, the ability to introduce a fully known stochastic mechanism in the design, under which exact inference can always be based, should not be sacrificed for the, often negligible, gain in efficiency achieved by choosing the best allocation(s), rather than choosing randomly from a smaller set of the nearly best allocations, as implied by a well-chosen rerandomization criterion.

As an illustration of the problem with standard asymptotic theory with rerandomization designs and the potential with the Fisher randomization test and correct asymptotics (Li et al., 2018), a small Monte Carlo study is conducted. We find that given correct inferential methods substantial gains in power can be made using rerandomization in comparison to complete randomization

REFERENCES

Cochran, W.G. & Cox, G.M. (1957) *Experimental designs*. Hoboken: John Wiley & Sons.

Imbens, G.W. & Rubin, D.B. (2015) *Causal inference in statistics, social, and biomedical sciences: An introduction*. Cambridge: Cambridge University Press.

Johansson, P. & Schultzberg, M. (2020) Rerandomization strategies for balancing covariates using pre-experimental longitudinal data. *Journal of Computational and Graphical Statistics*, 29, 798–813.

Kallus, N. (2018) Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 80, 85–112.

Kapelner, A., Krieger, A.M., Shalit, U. & Azriel, D. (2018) Harmonizing Fully Optimal Designs with Classic Randomization in Fixed Trial Experiments: 1–19. http://arxiv.org/abs/1810.08389.

Kapelner, A., Krieger, AM., Sklar, M. & Azriel, D. (2019) Optimal Rerandomization via a Criterion that Provides Insurance Against Failed Experiments: 1–27. http://arxiv.org/abs/1905.03337.

Krieger, A.M., Azriel, D. & Kapelner, A. (2019) Nearly random designs with greatly improved balance. *Biometrika*, 106, 695–701.

Lauretto, M.S., Stern, R.B., Morgan, K.L., Clark, M.H. & Stern, J.M. (2017) Haphazard intentional allocation and re-randomization to improve covariate balance in experiments. *AIP Conference Proceedings*, 1853.

Li, X., Ding, P. & Rubin, D. B. (2018) Asymptotic theory of rerandomization in treatment-control experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 9157–9162.

Morgan, K.L. & Rubin, D. B. (2012) Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40, 1263–1282.

Rubin, D.B. (1980) Randomization analysis of experimental data: The fisher randomization test. *Journal of the American Statistical Association*, 75, 575–582.

Schultzberg, M. & Johansson, P. (2020) Asymptotic inference for optimal rerandomization designs. *Open Statistics*, 1, 49–58. URL: https://www.degruyter.com/view/journals/stat/1/1/article-p49.xml.

Wu, C.-F. (1981) On the robustness and efficiency of some randomized designs. *The Annals of Statistics*, 9, 1168–1177.

How to cite this article: Johansson P, Rubin DB, Schultzberg M. On optimal rerandomization designs. *J R Stat Soc Series B*. 2021;83:395–403. https://doi.org/10.1111/rssb.12417