Journal of Statistical Planning and Inference 218 (2022) 43–58

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Rerandomization: A complement or substitute for stratification in randomized experiments?☆

Per Johansson, Mårten Schultzberg *

Department of Statistics, Uppsala university, Sweden

ARTICLE INFO

Article history:
Received 7 April 2020
Received in revised form 24 August 2021
Accepted 15 September 2021
Available online 24 September 2021

Keywords:
Blocked randomization
Experimental design
Mahalanobis distance
Rerandomization in tiers

ABSTRACT

Rerandomization is a strategy for improving balance on observed covariates in randomized controlled trials. It has been both advocated and advised against by renowned scholars of experimental design. However, the relationship and differences between stratification, rerandomization, and the combination of the two have not been previously investigated. In this paper, we show that stratified designs can be recreated by rerandomization and explain why, in most cases, stratification on binary covariates followed by rerandomization on continuous covariates is more efficient than rerandomization on all covariates at the same time.

1. Introduction

The most common design used to improve balance in a randomized control trials is stratified randomization, or blocked randomization. The idea to divide units into strata (i.e. groups/blocks) based on similarity on a set of covariates, and then to randomize the treatments within each stratum.

Another design is rerandomization. As the name suggests, rerandomization consists of redoing the randomization until some pre-specified balance criterion on the observed covariates (discrete or continuous) is met. That is, the randomization is restricted to a subset of allocations that fulfill a rerandomization covariate balance criterion. Rerandomization is computationally demanding compared to stratification. However, with modern computers this is no longer a real limitation.

Morgan and Rubin (2012) do not propose rerandomization as a substitute for stratification. Instead, the motivation for rerandomization is based on an understanding of that, also after blocking, randomization within strata can result in imbalances in other covariates. In this situation, Fisher is alleged to have recommended rerandomization (Morgan and Rubin, 2012). Athey and Imbens (2016) recommended researchers to first and foremost take care in the 'original design' to rule out unbalanced assignments instead of relying on rerandomization. This recommendation by the authors may be interpreted that they view rerandomization as a substitute for stratification which may be unfortunate if relevant continuous covariates are available. It is, however, arguably not obvious how or when to combine these strategies. This

☆ The authors thank Nikolay Angelov, Bength Muthén, Linda Muthén, Mattias Nordin and seminar participants at the Institute for Evaluation of Labour Market and Education Policy (IFAU) and the UppUpp conference, Uppsala University and SLU for helpful comments. Berndt Lundgren is acknowledged for kindly sharing data.

* Corresponding author.
E-mail address: marten.schultzberg@gmail.com (M. Schultzberg).

https://doi.org/10.1016/j.jspi.2021.09.002

paper contributes to the literature by comparing the properties and clarifying the relationship between stratification, rerandomization and the combination of both stratification and rerandomization.

Throughout the rest of this paper, we will use *stratified rerandomization* to mean to first stratify on a set of binary covariates and then rerandomize on remaining covariates, and *rerandomization* to mean to rerandomize on available covariates at the same time. The rerandomization makes use of the Mahalanobis distance between the means of the covariates of 'treated' and 'controls', together with an inclusion criterion for an allocation to be accepted. We give conditions for when the stratified design can be recreated by rerandomizing on the binary covariates used to form the strata. Utilizing this equivalence, we compare the relative efficiency of *stratified rerandomization* as compared to *rerandomization*. Given that the two criterions are minimized as a function of the sample size, N, the two designs are asymptotically equivalent with respect to efficiency. However, for moderate large N, *stratified rerandomization* is in general more efficient than *rerandomization*. With small N, the relationship may be reversed.

The efficiency comparison is studied using exact inference to units of the experiment. The main reason for this is clarity. The aim is to explain the relationship between stratification and rerandomization, and this is more clearly achieved for inference to the units of the experiments. Most of the theoretical results discussed in this paper extend to inferences to a population under random sampling. In addition, as will be shown, the choice of design is more complex in small sample settings where exact inference may be desirable for its lack of distributional assumptions.

The rest of this paper is structured as follows. Section 2 establishes the considered experimental designs. Section 3 compares the stratified design with the Mahalanobis-based rerandomization design and discusses the asymptotic efficiency of the designs under the Fisher null. Section 4 focuses on the relative efficiency of stratified rerandomization and rerandomization for inference to the units in the experiment. Section 5 presents a Monte Carlo study confirming the theoretical findings in Sections 3 and 4. Section 6 makes use of electricity consumption data as an illustration and Section 7 contains a discussion and concluding remarks.

2. Complete randomization, stratification and rerandomization

Let $Y_i(w)$ denote the potential outcome for unit i given the 'treatments' ($w = 0, 1$), e.g., treatment and control. For a sample of N experimental units, the sample average treatment effect is defined as

$$SATE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0)\right).$$

An experiment will have N_1 units assigned the treatment for which we observe $W_i = 1$, and N_0 units assigned the control for which $W_i = 0$ is observed. Under the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980) the observed Y_i is equal to $Y(W_i)$.

The differences-in-mean (DM) estimator is defined

$$\widehat{\tau} = \overline{Y}_1 - \overline{Y}_0, \tag{1}$$

where $\overline{Y}_w = \frac{1}{N_w}\sum_{i=1}^{N} W_i Y_i$.

In order to provide an intuition for the idea with experimental designs, let $\boldsymbol{W}$ be the $N \times \mathcal{N}_A$ matrix of all $\binom{N}{N_1} = \mathcal{N}_A$ possible allocation vectors under complete randomization. Furthermore let $\widehat{\tau}^j$ be the *estimate* for allocation j with assignment vector $\boldsymbol{W}^j \in \boldsymbol{W}$. The variance of the DM estimator can then be formulated as

$$V_{\boldsymbol{W}}(\widehat{\tau}) = \frac{1}{\mathcal{N}_A}\sum_{j=1}^{\mathcal{N}_A}(\widehat{\tau}_j - SATE)^2.$$

The idea of stratification and rerandomization is to remove allocations with potential large differences in $(\widehat{\tau}_j - SATE)^2, j = 1, \ldots, \mathcal{N}_A$. We let $\boldsymbol{W}^S \subset \boldsymbol{W}$, be the set of allocations under stratification and let $\boldsymbol{W}^\varphi \subset \boldsymbol{W}$ be the set of allocations under rerandomiziation. The cardinality of the two sets are denoted $\mathcal{N}_S$ and $\mathcal{N}_R$.

In the stratified design we form $s = 1, \ldots, S$ stratum, where each stratum is based on similarity on observed covariates. The stratified estimator is defined

$$\widehat{\tau}^{str} = \sum_{s=1}^{S}\frac{n^s}{N} \times \widehat{\tau}_s, \tag{2}$$

where $\widehat{\tau}_s$ is the mean differences estimator (1) in stratum s and n^s is the number of units in stratum s.

Rerandomization is more general than stratification in the sense that it can easily incorporate different types of covariates. The main difference is the exclusion criterion. With rerandomization the researcher first decide on a covariate balance measure, and then a criterion to exclude allocations that are not sufficiently balanced on the covariates. Given a random assignment that is sufficiently balanced, the analysis is in general based on the DM estimator (1).

2.1. Mahalanobis-based rerandomization

Due to the well known properties of the Mahalanobis distance, we restrict the analyses to Mahalanobis-based rerandomization designs and discuss the most important results from Morgan and Rubin (2012).[1]

Let $\boldsymbol{X}$ be the $N \times K_0$ matrix of fixed covariates, and, for simplicity, let $N_1 = N_0$. For a given allocation j the Mahalanobis distance between the covariate mean vectors of the units assigned to treatment and control, respectively, is defined as

$$M(\boldsymbol{X}, \boldsymbol{W}^j) = \frac{N}{4}(\overline{\boldsymbol{X}}_1^j - \overline{\boldsymbol{X}}_0^j)' cov(\boldsymbol{X})^{-1}(\overline{\boldsymbol{X}}_1^j - \overline{\boldsymbol{X}}_0^j),\ j = 1, \ldots, \mathcal{N}_A, \tag{3}$$

where $\overline{\boldsymbol{X}}_1^j - \overline{\boldsymbol{X}}_0^j$ is the difference in mean vectors of treated ($\overline{\boldsymbol{X}}_1^j$) and controls ($\overline{\boldsymbol{X}}_0^j$), which is a $K_0 \times 1$ stochastic vector as it depends on the random allocation.

Morgan and Rubin (2012) suggested accepting the treatment assignment vector $\boldsymbol{W}^j$ only when

$$M(\boldsymbol{X}, \boldsymbol{W}^j) \leq a_0,$$

where a_0 is a positive constant called the *rerandomization criterion*. This means that the set $\boldsymbol{W}^{\varphi}$ is implicitly defined as $\forall \boldsymbol{W}^j \exists M(\boldsymbol{X}, \boldsymbol{W}^j) \leq a_0$

If the covariate means are normally distributed, $M(\boldsymbol{X}, \boldsymbol{W}^j) \sim \chi^2(K_0)$. This implies that a_0 can be indirectly determined by setting

$$p_{a_0} = \Pr(\chi^2(K_0) \leq a_0),$$

and that allocations can be sampled from any desired percentile of allocations with the Mahalanobis distances smaller or equal to a_0. As the number of randomizations needed to draw an allocation that fulfills the criterion is geometrically distributed with expected value $1/p_{a_0}$, the expected number of randomizations before drawing an allocation fulfilling the criterion with, e.g., $p_{a_0} = 0.001$, is 1000.

Morgan and Rubin (2015) extend this idea by proposing the rerandomization to be done in tiers of covariates. The most important covariates should be placed in the first tier, often using a strict rerandomization criterion, and the second most important covariates in tier two with a slightly less restrictive criterion, etc. All allocations with large imbalances in the covariates of the first tier are excluded. In the second tier only the admissible allocations in the first tier are considered etc. This means that the number of possible allocations decreases for each tier, until only allocations fulfilling the overall balance criteria remain in the final tier.

By placing all categorical covariates in the first tier and all continuous covariates in the second tier, rerandomization in tiers can be seen as special case of stratified rerandomization. The main difference is that rerandomization in tiers allows for any type of covariate in the first tier and has an explicit rerandomization criterion for this tier, making it possible to put covariates in tiers based on their believed relative importance rather than variable type.

3. A comparison of stratification with rerandomization

To facilitate the understanding of the relative efficiency of the different designs, we restrict the comparison to an additive treatment effect, or the Fisher null, i.e., $H_0^{Fisher} : Y_i(1) - Y_i(0) = 0\ \forall\ i = 1, \ldots, N$. The variance reduction of the treatment effect will be larger if the effect is heterogeneous with respect to the included covariates for both stratification (see e.g. Imbens and Rubin, 2015) and with rerandomization (see Li, Ding and Rubin 2018). For this reason, we do not view this restriction as important for the comparison of the relative efficiency also under asymptotic inferences under the Neyman null.

Let $\boldsymbol{X}_1$ be a $N \times K_1$ matrix of binary covariates and let $\boldsymbol{X}$ be the $N \times K_0$ matrix containing the K_1 binary covariates in $\boldsymbol{X}_1$ and all their interactions, implying $K_0 = \sum_{i=1}^{K_1} \binom{K_1}{i}$. Now, consider the linear projection of the outcome on the covariates, that is,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$.

Theorem 3.1. *Let $\boldsymbol{W}^o$ be the set of allocations minimizing the conditional variance of the outcome under the sharp null. Then this set can be obtained as*

$$\boldsymbol{W}^o = \min_{\boldsymbol{W}^j \in \boldsymbol{W}} M(\boldsymbol{X}, \boldsymbol{W}^j)$$

Proof. As $\boldsymbol{X}$ contains only binary covariates including all interactions, the linear projection of $\boldsymbol{Y}$ on $\boldsymbol{X}$ (cf. Eq. (4)) is fully saturated and is therefore equal to the conditional expectation of $\boldsymbol{Y}$ under the sharp null. This, together with the fact that the Mahalanobis distance is affinely invariant, implies that minimizing the variance in $\boldsymbol{X}$ will directly minimize $Var(\boldsymbol{Y}|\boldsymbol{X})$ for all $\boldsymbol{\gamma}$ in Eq. (4). □

[1] See Johansson and Schultzberg (2020) for an alternative rerandomization design.

Theorem 3.1 implies that rerandomizing on binary covariates and all their interactions, using the strictest possible Mahalanobis criterion, always randomizes within the set of allocations with the minimum variance in the outcome under the sharp null.

Consider now the Mahalanobis-based randomization based on the $N \times S$ matrix $\boldsymbol{D} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_S)$ where $\boldsymbol{D}_s \; \forall \; s = 1, \ldots, S$, are $N \times 1$ vectors with one for units belonging to stratum s and zero otherwise. Thus $\boldsymbol{N}^S = \boldsymbol{D}'\boldsymbol{1}_N = (n^1, n^2, \ldots, n^S)'$ is the $S \times 1$ vector of the number of units in each stratum. Define

$$\boldsymbol{W}^{\varphi} = \min_{\boldsymbol{W}^j \in \boldsymbol{W}} M(\boldsymbol{D}, \boldsymbol{W}^j)$$

Given that $N_1 = N_0$ the mean-difference vector of treated and control group for allocation j is

$$\overline{\boldsymbol{D}}_1^j - \overline{\boldsymbol{D}}_0^j = \frac{\boldsymbol{N}_{1j}^S}{N_1} - \frac{\boldsymbol{N}_{0j}^S}{N_0} = \frac{2\boldsymbol{N}_{1j}^S - \boldsymbol{N}^S}{N/2},$$

where $\boldsymbol{N}_{1j}^S = (n_{1j}^1, n_{1j}^2, \ldots, n_{1j}^S)'$ and $\boldsymbol{N}_{0j}^S = (n_{0j}^1, n_{0j}^2, \ldots, n_{0j}^S)'$ are the vectors of number of treated and controls in each stratum for allocation j, respectively. This means that $\min_{\boldsymbol{W}^j \in \boldsymbol{W}} M(\boldsymbol{D}, \boldsymbol{W}^j) := 0$ for all j where $2\boldsymbol{N}_{1j}^S = \boldsymbol{N}^S$. As $n^S = N - \sum_{s=1}^{S-1} n^s$ this means that we can drop the last column and let $\boldsymbol{X} = (\boldsymbol{1}_N, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_{S-1})$, where $\boldsymbol{1}_N$ is a column vector with ones.

Corollary 3.1. *If $n^s \bmod 2 = 0 \; \forall \; s = 1, \ldots, S$, it follows that*

$$\boldsymbol{W}^o = \boldsymbol{W}^S = \boldsymbol{W}^{\varphi} : M(\boldsymbol{X}, \boldsymbol{W}^j) := 0.$$

Proof. Since the Mahalanobis distance is affinely invariant, it holds that

$$M(\boldsymbol{D}, \boldsymbol{W}^j) := M(\boldsymbol{X}, \boldsymbol{W}^j),$$

and the proof follows directly from Theorem 3.1. □

Corollary 3.1 shows that, in the case when all strata are of even sample size, stratification gives exactly the same design as rerandomization with the rerandomization criterion zero.[2] In other words, letting $a_0 = 0$ the Mahalanobis-based randomization will by design try to find a design that are balanced within each strata. This is however only possible if all strata are of even sample size, that is, $n^s \bmod 2 = 0$, for all s.

When $n^{s'} \bmod 2 \neq 0$ for $s' \in (1, \ldots, S)$, and with the aim of letting $N_1 = N_0$, the researcher would randomly assign $n^{s'}/2 - \frac{1}{2}$ or $n_1^{s'}/2 + \frac{1}{2}$ to be treated. In the first case $n_1^{s'} - n_0^{s'} = -1$ and in the second $n_1^{s'} - n_0^{s'} = 1$. In this situation $\boldsymbol{W}^S$ cannot be shown to be equal to $\boldsymbol{W}^o$.

Restricting randomization to $\boldsymbol{W}^S$ removes imbalances within each stratum but does not guarantee that the covariates are balanced over the full sample. This is not a problem for the DM estimator (cf Eq. (2)) as the within strata estimators are all unbiased. With the set $\boldsymbol{W}^o$ we are guaranteed to obtain an estimator with an overall minimum variance, under the null, which one is more efficient will depend on the context.

3.1. Relative efficiency

Before consider the efficiency and computational time with a fixed N it is useful to consider the asymptotic efficiency under the Fisher null. Let R^2 is the squared multiple correlation between $\boldsymbol{Y}$ and $\boldsymbol{X}$, or the coefficient of determination in (4).

3.1.1. Stratification

Imbens and Rubin (2015, p. 206) show that the OLS estimator of (4) converges to $\tau_w = \sum_{s=1}^{S} \omega_s \tau_s / \sum_{s=1}^{S} \omega_s$, where τ_s is the treatment effect in stratum s. In our setting this means that $\omega_s = \frac{n^s}{N} \times (\frac{n_1^s}{n^s}(1 - \frac{n_1^s}{n^s}))$.

In a balanced design within each stratum, i.e. $n_1^s = n^s/2 \; \forall s$, $\omega_s = \frac{n^s}{N} \times 0.25$. This means that the OLS estimator is asymptotically equivalent to the stratified estimator in this specific situation.[3]

Let $\widehat{\tau}^{cr}$ be the DM estimator under complete randomization. Given that $n_1^s = n^s/2 \forall s$, and that the covariates by design is independent of the error term, we get $Var(\widehat{\tau}^{str}) = (1 - R^2)Var(\widehat{\tau}^{cr})$. Here, $Var(\widehat{\tau}^{cr}) = \sigma_Y^2(1/N_1 + 1/N_0)$, where σ_Y^2 is either, the within sample variance of $\boldsymbol{Y}$, or the population variance. The percent reduction in sampling variance (PRIV) of the treatment effect under stratification against complete randomization is then

$$PRIV_S = 100 \times \frac{Var(\widehat{\tau}^{cr}) - Var(\widehat{\tau}^{str})}{Var(\widehat{\tau}^{cr})} = 100 \times R^2. \tag{5}$$

[2] Note that this equivalence is also true with heterogeneous treatment effects, thus the fact that we restrict the comparison to the situation with an additive effect is no restriction when all strata are of even size.

[3] Imbens and Rubin (2015, p. 206) also show that with an augmented regression model, including interaction terms with treatment and stratum, the OLS estimator is asymptotically equivalent to the stratified estimator also in unbalanced designs.

3.1.2. Mahalanobis-based rerandomization

Here $\mathbf{X}$ is a $1 \times K_0$ vector of covariates. Under the assumption that errors terms in the regression model (4) are normally distributed,[4] Morgan and Rubin (2012) show that PRIV of the treatment effect under Mahalanobis-based rerandomization against complete randomization is equal to

$$PRIV_1 = 100 \times \frac{Var(\widehat{\tau}^{cr}) - Var(\widehat{\tau}^{rr})}{Var(\widehat{\tau}^{cr})} = 100 \times R^2(1 - \nu_0), 0 \leq \nu_0 \leq 1 \tag{6}$$

where

$$\nu_0 = \frac{\Pr(\chi^2(K_0 + 2) \leq a_0)}{\Pr(\chi^2(K_0) \leq a_0)}, \tag{7}$$

and $\widehat{\tau}^{rr}$ is the DM estimator (1) under Mahalanobis-based rerandomization.

As the probability of a Chi-square distributed variable to be less than a_0 is decreasing with the degrees of freedom, $\nu_0 \to 0$ as $a_0 \to 0$. The implication is, thus, that with a rerandomization criterion a_0 close to zero, such that $\nu_0 \simeq 0$, stratification and rerandomization should give similar improvements in efficiency.

With T tiers with covariates $\boldsymbol{X}_t$, for $t = 1, \ldots, T$, such that $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T)$ and $\boldsymbol{X}_t$ an $N \times K_t$ matrix, Morgan and Rubin (2015) show that

$$PRIV_T = (1 - \nu_1)R^2_{1:1,y} + \sum_{t=2}^{T}(1 - \nu_t)(R_{1:t,y} - R_{1:(t-1),y}), \tag{8}$$

where $R^2_{1:t,y}$ is squared multiple correlation between $\boldsymbol{Y}$ and $\boldsymbol{X}_{1:t} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_t)$ and a_t, is the rerandomization criterion for tier t, implying

$$\nu_t = \frac{\Pr(\chi^2(K_t + 2) \leq a_t)}{\Pr(\chi^2(K_t) \leq a_t)}, t = 1, \ldots, T. \tag{9}$$

4. Stratified rerandomization and rerandomization

Since Mahalanobis-based rerandomization on categorical covariates with interactions can be made equivalent to stratification, it follows that the relative performance of stratified rerandomization and rerandomization can be investigated using the framework of rerandomization in tiers. As strict equivalence is possible only under Corollary 3.1; we restrict the comparison to the completely balanced experiment accordingly.

Let $\boldsymbol{X}_1 = (\mathbf{1}_N, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_{S-1})$, $\boldsymbol{X}_2$ be the set of K_2 continuous covariates, and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ be the set of K_0 covariates. The PRIV for the DM estimator under rerandomization and rerandomization in tiers (used for stratified rerandomization) are given in Eqs. (6) and (8). In this setting we get for $t = 1$ and 2

$$PRIV_1 = 100 \times (1 - \nu_0)R^2$$

and

$$PRIV_2 = (1 - \nu_1)R^2_1 + (1 - \nu_2)(R^2 - R^2_1),$$

respectively. Here R^2_1 is squared multiple correlation between $\boldsymbol{Y}$ and $\boldsymbol{X}_1$.

Due to the balanced strata, rerandomizing on $\boldsymbol{X}_1$ with the zero criteria (stratifying) implies $a_1 = p_{a_1} = \nu_1 = 0$. That is, for stratified rerandomization, all the variance in $\widehat{\tau}$ from $\boldsymbol{X}_1$ is controlled for and the number of remaining allocations is $\mathcal{N}_S$.

As a tool for comparing the two designs, we use the ratio of the PRIV's of the two designs

$$\begin{aligned} RPRIV &= \frac{PRIV_2}{PRIV_1} \\ &= \frac{(1 - \nu_2)R^2 + \nu_2 R^2_1}{(1 - \nu_0)R^2} \\ &= \frac{1 - \nu_2}{1 - \nu_0} + \frac{\nu_2}{1 - \nu_0}\frac{R^2_1}{R^2}. \end{aligned} \tag{10}$$

It is reasonable to let a_0 and a_2 tend to zero with N. Then for any R^2_1, $R^2 > 0$

$$\frac{1 - \nu_2}{1 - \nu_0} + \frac{\nu_2}{1 - \nu_0}\frac{R^2_1}{R^2} \xrightarrow{p} 1 \text{ as } N \to \infty. \tag{11}$$

[4] The errors are by definition uncorrelated with $\boldsymbol{X}$. The assumption is needed as the proof requires independence. When $\boldsymbol{X}$ consists of a set of binary covariates and all their interactions the normality assumption can be relaxed as then Eq. (4) is the conditional mean.

This follows from the fact that both ν_0 and ν_2 converge to zero with N as a_0 and a_2 tend to zero with N. Thus, the two designs are given equal variance reductions asymptotically.

To get an understanding of the behavior and limitations of the two designs, we evaluate the relative efficiency of the two procedures, for the more realistic situation with N fixed in the next sub-section.

4.1. Efficiency and computational time with fixed n

Restricting the focus of inference to the units of the sample, all stochastic variation in an experiment stems from the randomization, which means that the lowest level of risk in a FRT is determined by the number of possible treatment allocations in the randomized experiment. This means that if the lowest risk of making a false decision is chosen to be 5%, the number of possible allocations in a two-sided test needs to be at least 40. For 1%, the corresponding number of allocations is 200. That is, the lowest possible risk in a double sided exact test can be viewed as the resolution, r, of the exact p-value in the FRT. With e.g. $\mathcal{N}_R$ possible allocations in a rerandomization design, $r = 2/\mathcal{N}_R$.

For most sample sizes, $\mathcal{N}_R$ is large for any rerandomization criteria. Therefore, a too large r is not usually an issue. A too large number of allocations, is on the other hand, a problem as it makes it intractable to calculate the exact p-value. This issue was pointed out by Athey and Imbens (2016), and used as an argument for recommending researchers to *not* use rerandomization. The rapid growth of combinations is, however, a potential problem for using exact tests under any randomized design, not only rerandomization. For example, in pairwise stratification (each stratum is of size two), which implies

$$\mathcal{N}_S = \binom{2}{1}^{N/2}.$$

Already for $N = 60$, there are 1.0737×10^9 possible allocations which is impossible to manage with a typical computer. One common solution is to approximate the exact p-value by Monte Carlo simulation. An alternative, suggested in Johansson and Schultzberg (2020), is to do an exact test on a limited set of allocations of 'optimal' allocations found by rerandomization.

Before discussing this procedure, it is useful to first provide the intuition for the unbiasedness of the estimators in the sets $\boldsymbol{W}, \boldsymbol{W}^S$ or $\boldsymbol{W}^\varphi$. For a given random allocation $\boldsymbol{W}^j$ with an estimate $\widehat{\tau}^j$ the estimate of the 'mirror allocation' $\boldsymbol{W}^{j'} = \mathbf{1} - \boldsymbol{W}^j$ is simply $-\widehat{\tau}^j$. Thus, any set $\boldsymbol{W}^\varphi$ with cardinality two or larger containing only mirror allocations will be unbiased. As the Mahalanobis distance is affinely invariant it is also the case that $M(\boldsymbol{X}, \boldsymbol{W}^j) \equiv M(\boldsymbol{X}, \mathbf{1} - \boldsymbol{W}^j)$. This symmetry, thus, provides an intuition for why the DM estimator is unbiased under Mahalanobis-based rerandomization. The implication for any algorithm with an aim of finding a set of allocations with cardinality $\mathcal{H}$ is that, by only sampling allocations from the first half of the lexicographically ordered allocations, only $\mathcal{H}/2$ allocations fulfilling the criterion must be found after which the corresponding mirror allocations are added to the set.

The implied smallest number of allocations needed for inference is then $\mathcal{H} = 2/r$. The set $\boldsymbol{W}^{\varphi*}$ is the optimal set of allocations, conditional on the Mahalanobis distance balance measure, if it contains the allocations with the $\mathcal{H}/2$ first unique order statistics of the Mahalanobis distance across all $\mathcal{N}_A$ allocations as we include also the mirror allocations. To be clear, the optimal set of allocations is obtained in $(M^{[1]}, \ldots, M^{[H/2]})$, where $M^{[j]}, j = 1, \ldots, N_A/2$ denotes the order statistics of the Mahalanobis distance statistics. For normally distributed covariates or large sample settings, this corresponds to using the rerandomization criterion $a_0 = M^{[H]}$ in $p_{a_0} = \Pr(\chi^2(K_0) \leq a_0)$, which implies that $\mathcal{H} = \mathcal{N}_A p_{a_0}$.

When N is large, the Mahalanobis distance cannot within a reasonable time limits be calculated over all of the $\mathcal{N}_A/2$ allocations due to the rapid growth of $\binom{N}{N_1}$. In this situation, the algorithm suggested in Johansson and Schultzberg (2020) sequentially keeps the $\mathcal{H}/2$ allocations with the smallest Mahalanobis distance over subsets until a total number of $\mathcal{I}$ allocations from the original $\mathcal{N}_A/2$ has been drawn. When the algorithm is finished, the mirror allocations are included to give $\mathcal{H}$ in total. The final $\mathcal{H}$ allocations with the smallest Mahalanobis distances are then used as $\boldsymbol{W}^{\varphi*}$. This procedure differs from the procedure suggested in Morgan and Rubin (2012) in the sense that p_{a_0} is not set before the rerandomization. Instead p_{a_0} is a function of $\mathcal{I}$ and $\mathcal{H}$. The only restriction is computational time and the implied a_0 can be probabilistically bounded by setting $\mathcal{I}$ accordingly (Johansson and Schultzberg, 2020). Within each set $\mathcal{I}$, the Mahalanobis distances are Chi square-distributed, with the implication that the probability for an allocation to be accepted will depend on $\mathcal{I}$, that is $p^{\mathcal{I}}_{a_0} = \Pr(\chi^2(K_0) \leq a_0 | \mathcal{I})$ where $\lim_{\mathcal{I} \to \mathcal{N}_A} p^{\mathcal{I}}_{a_0} = p_{a_0}$. In expectation $p^{\mathcal{I}}_{a_0} = \mathcal{H}/\mathcal{I}$.

4.1.1. The relative efficiency for fixed N

Given a level of resolution r in the exact p-value of a two-sided hypothesis test, the minimum number of allocations that must remain after stratification is $\mathcal{H} = 2/r$. Under the assumption

$$M(\boldsymbol{X}_2, \boldsymbol{W}^j) \sim \chi^2(K_2),$$

it follows that $\mathcal{H} = \mathcal{N}_S p_{a_2}$, which implies that rerandomization criterion in the second tier of the stratified rerandomization is bounded

$$\frac{\mathcal{H}}{\mathcal{N}_S} \leq p_{a_2} \leq 1. \tag{12}$$

Thus, when $\mathcal{N}_S \simeq \mathcal{H}$, $p_{a_2} = \nu_{t_2} \simeq 1$. Putting it differently, if only a few allocations can be discarded under the rerandomization based on the second tier, one cannot expect substantial variance reductions in the covariates in the second tier. This suggests that, if $\mathcal{N}_S$ is close to $\mathcal{H}$, and $\boldsymbol{X}_2$ is believed to explain a lot of variation in the outcome, it might be a bad idea to first stratify.

In order for the stratified rerandomization to have larger percentage reduction in variance than rerandomization, it must hold that

$$\frac{1-\nu_2}{1-\nu_0} + \frac{(R_1^2/R^2)\nu_2}{1-\nu_0} > 1, 0 \leq \nu_0, \nu_2 \leq 1.$$
$$\iff$$
$$(1 - \frac{R_1^2}{R^2}) < \frac{\nu_0}{\nu_2}, 0 \leq \nu_0, \nu_2 \leq 1. \tag{13}$$

As $0 \leq R_1^2/R^2 \leq 1$, this means that whenever $\frac{\nu_0}{\nu_2} > 1$, stratified rerandomization is more efficient than rerandomization in expectation.

Let $Q(r, K)$ be the quantile function of the Chi-square distribution such that $Q(p_{a_t}, K_t) = a_t$, for $t = 0, 1, 2$. For a given r, the optimal criteria for the stratified rerandomization and rerandomization are $a_2 = Q(r, K_2)$ and $a_0 = Q(r, K_0)$, respectively. With $r = \mathcal{H}/\mathcal{N}_S$ and $\mathcal{H}/\mathcal{N}_A$ for stratified rerandomization and rerandomization, respectively, we get

$$\nu_0/\nu_2 = \frac{\Pr\left(\chi^2(K_0+2) \leq Q(\frac{\mathcal{H}}{\mathcal{N}_A}, K_0)\right)}{\Pr\left(\chi^2(K_0) \leq Q(\frac{\mathcal{H}}{\mathcal{N}_A}, K_0)\right)} \Big/ \frac{\Pr\left(\chi^2(K_2+2) \leq Q(\frac{\mathcal{H}}{\mathcal{N}_S}, K_2)\right)}{\Pr\left(\chi^2(K_2) \leq Q(\frac{\mathcal{H}}{\mathcal{N}_S}, K_2)\right)}. \tag{14}$$

From this expression it becomes clear that the relative efficiency of the DM estimator under stratified rerandomization and rerandomization depends on the degrees of freedom in the Chi-square distribution of the Mahalanobis distances and the number of possible allocations in the rerandomization step of the two designs (i.e. $\mathcal{N}_A$ and $\mathcal{N}_S$). The efficiency of rerandomization is decreasing in the degrees of freedom and increasing in the number of allocations. The stratification reduces both the degrees of freedom (from K_0 to K_2) and the number of allocations (from $\mathcal{N}_A$ to $\mathcal{N}_S$) in the second tier rerandomization.

Variance-reduction evaluation for pairwise stratification. To illustrate how the relative efficiency of stratified rerandomization compared to rerandomization depends on R_1^2/R^2, we consider the case with $N/2$ strata of size 2 for $N = 12, 14, 16, 18, 20, 22, 24$, and one continuous covariate. With $K_1 = N/2 - 1$ binary and one continuous covariate the total number of covariates is $K_0 = N/2$. We vary the lowest level of risk to be 5%,1% and 0.1%, which means that $\mathcal{H} = 40{,}200$ and 2000.

From Fig. 1 we can see that with a 5% level of risk, rerandomization is preferable to stratified rerandomization, i.e. has larger expected PRIV, only when $R_1^2/R^2 \leq 0.20$ and $N = 12$. When the level of risk is set to 1%, $\mathcal{N}_S < \mathcal{H}$ for $N \leq 14$, which implies that stratified rerandomization is not an option. For $N = 16$, rerandomization is preferable to stratified rerandomization when $R_1^2/R^2 \leq 0.53$. For larger experiments, stratified rerandomization is preferable for all R_1^2/R^2. With the level of risk set to 0.1%, $\mathcal{N}_S < \mathcal{H}$ for $N \leq 20$, which means that stratified rerandomization is not an option in these cases. For $N = 22$, rerandomization is preferable to stratified rerandomization when $R_1^2/R^2 \leq 0.78$, and for $N = 24$, stratified rerandomization is preferable for all R_1^2/R^2.

Fig. 1 clearly shows the trade off between the number of remaining allocations after stratification and the 'cost' of increasing the degrees of freedom in the Chi-square distribution of the Mahalanobis distance for the rerandomization. By first stratifying, the number of degrees of freedom in the Chi-square distribution of the Mahalanobis distances in the remaining rerandomization is reduced from $N/2$ to one. Lower degrees of freedom means less diffused Mahalanobis distances which provides better precision in the rerandomization in the second tier. On the other hand, if there are few allocations left after the stratification, the rerandomization on the remaining covariates becomes restricted, as $\mathcal{H}$ allocation must be kept for inference based on the choice of the maximum level of risk. It is important to understand that the pairwise stratification is a 'worst case' scenario for rerandomization as the difference between K_0 and K_2 is maximized for each N in this design. If the number of binary covariates is fixed over N, stratified rerandomization and rerandomization would give more similar PRIV in accordance with Eq. (11).

If there is no a priori information on the relative importance of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ in explaining the outcome, it is reasonable to assume that $\boldsymbol{X}_1$ explains as much as $\boldsymbol{X}_2$, i.e., $R_1^2/R^2 = 0.5$. With $R_1^2/R^2 = 0.5$ the stratified rerandomization is in expectation more efficient than Mahalanobis-based rerandomization on $\boldsymbol{X}$ whenever $\nu_0/\nu_2 > 0.5$ (see Eq. (13)). For $N = 16$ we saw that exact inference is possible under both stratified rerandomization and rerandomization with $\alpha = 1\%$. As rerandomization was more efficient for all $R_1^2/R^2 \leq 0.53$, this implies that rerandomization is preferable with an agnostic assumption on the importance of the two types of covariates dependence with the outcome in this case. However, as these results build on asymptotic properties they should be interpreted with caution. Finite sample properties in this case will be presented in Section 5.

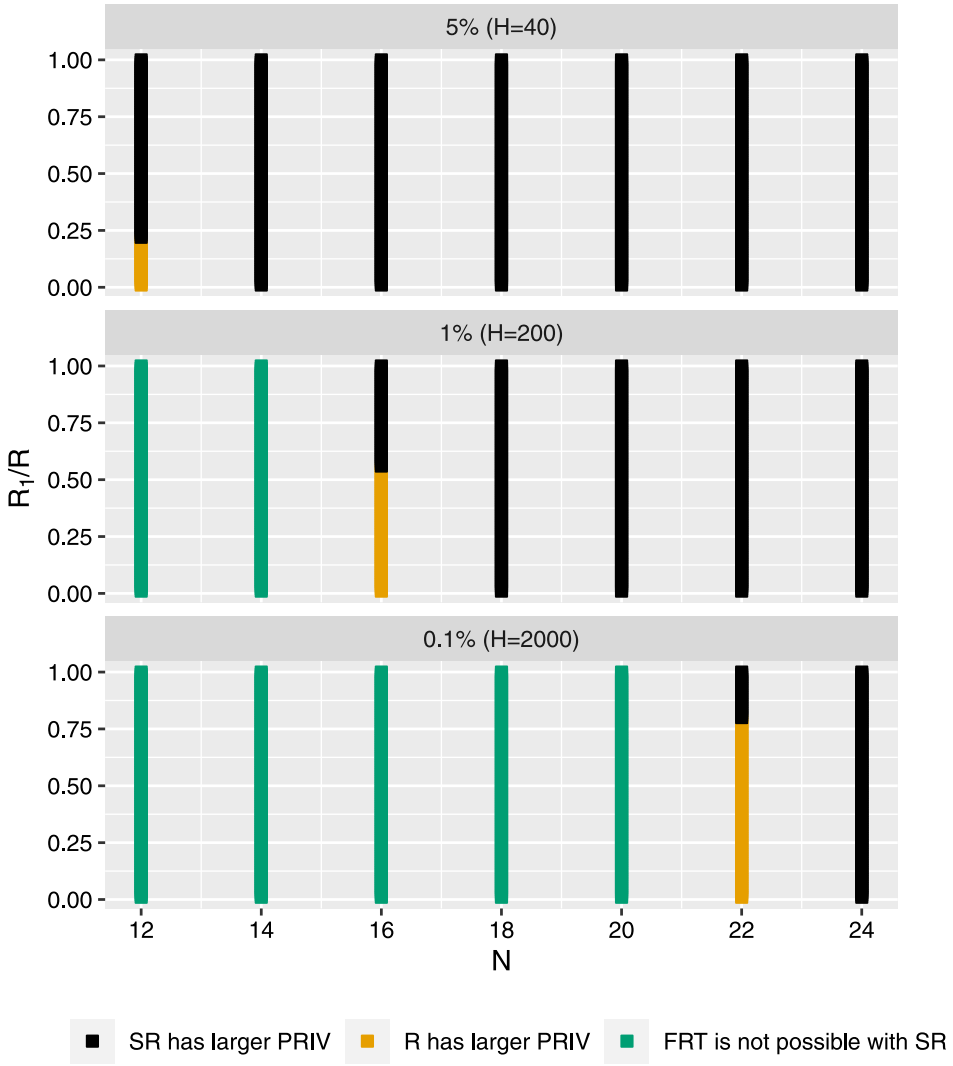

Fig. 1. Comparison of expected PRIV under optimal stratified rerandomization (SR) and rerandomization (R) designs for sample sizes between 12 and 22, for a test of level 5%, 1% and 0.1%.

4.1.2. Computational time as a function of the number of covariates

In the previous sub-section it was shown that, for sample sizes smaller than 24, the number of remaining allocations after stratification on $\boldsymbol{X}_1$ restricts the subsequent rerandomizations on $\boldsymbol{X}_2$ so much that rerandomization on $\boldsymbol{X}$ may be preferable. However, this restriction is only a concern for these small experiments. Even with moderately large experiments, it is intractable to go through all allocations (also after stratification) wherefore it is important to take the computational time it takes to find $\mathcal{H}$ acceptable allocations into account when comparing designs using rerandomization.

Here we study the computational time for cases where the sample size is too large for exhaustive rerandomization and the distribution of the mahalanobis distance is well approximated by a Chi-square distribution to understand how stratified randomization can improve the design as compared to rerandomization, given a fixed time budget for the design.

The expected number of considered allocations needed to find one acceptable allocation for any given criterion a_t in tier t using rerandomization is $1/p_{a_t}$. This means that, on average, $\mathcal{H}/p_{a_t}$ allocations need to be sampled to obtain $\mathcal{H}$ acceptable allocations. For $\mathcal{H} = 40$, with $p_{a_t} = 0.00001$, 4 million allocations needs to be sampled to obtain 40 allocations that fulfills the criterion on average. However, it is ν_t and not p_{a_t} that determines the efficiency gain from the design as, for a given R^2, the PRIV is only a function of ν_t. As ν_t increases with K_t, the variance reduction from the rerandomization decreases in K_t for a fixed p_{a_t}. This means that in order to achieve the same variance reduction from a large set as for a small set of covariates, the criterion p_{a_t} needs to be reduced, and, therefore the number of sampled allocations needs to be increased. An alternative to searching for the $\mathcal{H}$ optimal allocations among all $\mathcal{N}_A$ allocations we used the algorithm

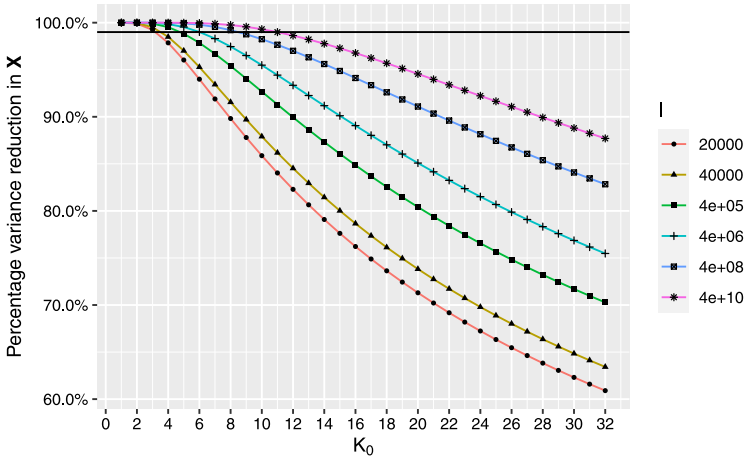

Fig. 2. Expected percentage variance reduction in the rerandomization covariates for various numbers of considered allocations as a function of degrees of freedom in the rerandomization. Values above the solid line imply that more than 99% of the variation in the covariates is removed by the rerandomization.

Table 1
Estimated time consumption (years) for calculating the Mahalanobis distance for all $\binom{N}{N/2}$ allocations, assuming that 1,000,000 allocations can be considered each second.

N	50	60	70	80	90	100
Expected time (years)	4	3750	3.56×10^6	3.41×10^9	3.29×10^{12}	3.20×10^{15}

suggested in Johansson and Schultzberg (2020). Thus, we let $p^{\mathcal{I}}_{a_t} = \Pr(\chi^2(K_t) \leq a_t|\mathcal{I})$, where $\mathcal{I}$ is the number of allocations in a randomly drawn subset of $\boldsymbol{W}$, $\boldsymbol{W}^S$ or $\boldsymbol{W}^\varphi$.

We study the computational time for a fixed $\mathcal{H}$ in an experiment where $\mathcal{I}$ is varied between 20 thousand and 40 billions for $K_0 = 1, \ldots, 32$, in other words, at different levels of $p^{\mathcal{I}}_{a_0}$. Here N is fixed an assumed large enough for the Mahalanobis distance to be well approximated by a Chi-square distribution and too large for going through all allocations. The results of this exercise are displayed in Fig. 2.

Fig. 2 shows the expected PRIV from rerandomization. It is apparent from the figure that the relation between the expected variance reduction and the number of sampled allocations, $\mathcal{I}$, is non-linear in K_0. For $K_0 \leq 3$ the number of considered allocations that remove all of the variation is small. For K_0 larger than five, the number of considered allocations needed to reduce all the variations becomes implausibly large. For $K_0 \geq 11$, not even 40 billion allocations is enough to get 99% reduction in PRIV. This illustrates the potential benefits of reducing the number of degrees of freedom in the second tier by using stratified rerandomization, and why rerandomizing in tiers is a good idea in general.

These results are of importance for the comparison of stratified rerandomization and rerandomization. That is, since $K_2 < K_0$ by construction, the computational time of the rerandomization step in stratified rerandomization may be substantially smaller than in rerandomization, especially if K_1 is large.

A perspective on computational time. The seemingly large numbers of considered allocations in Fig. 2 are in fact a very small part of the total number of allocations for moderately large samples. For example, if the sample size is 50, 4×10^{10} constitutes $100(4 \times 10^{10}/\binom{50}{25}) = 0.03\%$ of all possible allocations.

The implication of the growth of combinations is difficult to comprehend but to give a perspective, we exemplify by predicting the computational time for finding the globally best allocation for $N = 50, 60, 70, 80, 90, 100$. A decently fast software implementation can go through around 1,000,000 allocations per second[5] depending on the sample size and the number of covariates. Table 1 displays the estimated computational time for calculating all the Mahalanobis distances. Already for $N = 50$ it takes 4 years, and for $N = 60$ the computational time is more than 3000 years, indicating the complexity of finding the allocations with the $\mathcal{H}$ globally smallest Mahalanobis distances. This problem can be fully parallelized, and there are likely software implementation that can speed up the calculations by some factor. However, given the rapid increase, going through all allocations for samples sizes such as $N = 100$ is still completely intractable with current hardware and software.

[5] The figure 1,000,000 comes from timing an implementation in the programming language Julia v1.1.0., with one covariate rounding up the number of consider allocations. The corresponding figure for an implementation in base R v3.5.3 is around 30,000.

5. Monte Carlo simulations

Three Monte Carlo (MC) simulations are conducted to study the power of the DM estimator in an exact FRT under stratification (S), stratified rerandomization (SR), Mahalanobis-based rerandomization (MR_0), and complete randomization (C).

Data are generated as

$$Y_i(0) = \boldsymbol{x}_{1i}\boldsymbol{\beta}_1 + \beta_2 x_{2i} + \epsilon_i \tag{15}$$

where $\boldsymbol{x}_{1i}$ is the $K_1 \times 1$ vector of binary covariates including their interactions, and x_{2i} and ϵ_i are both i.i.d. exponentially distributed variables: $x_{2i} \sim \exp(\lambda_1)$ and $\epsilon_i \sim \exp(\lambda_2)$. The alternative for which the power is studied is set to $Y_i(1) = Y_i(0) + 0.6$. The parameters are chosen such that in expectation $R^2 = 50\%$ in all settings. The number of replications in each cell is 5000 and the exact p-value is defined as

$$\pi_{mj} = \frac{1}{\mathcal{H}} \sum_{r=1}^{\mathcal{H}} \mathbf{1}(|\widehat{\tau}(\boldsymbol{W}_r^*, \boldsymbol{Y}_j)| \geq |\widehat{\tau}^j|), j = 1, \ldots, \mathcal{H},\ m = 1, \ldots, 5,000, \tag{16}$$

where $\widehat{\tau}(\boldsymbol{W}_r^*, \boldsymbol{Y}_j)$ is the distribution of estimates over all allocations r given allocation j (for any of the sets $\boldsymbol{W}^* = \boldsymbol{W}$, $\boldsymbol{W}^* = \boldsymbol{W}^S$ or $\boldsymbol{W}^* = \boldsymbol{W}^{\varphi*}$) in replication m. The power in replicate m is calculated as

$$P_m = \frac{1}{\mathcal{H}} \sum_{j=1}^{\mathcal{H}} \mathbf{1}(\pi_{mj} \leq 0.05). \tag{17}$$

The designs of the first two Monte Carlo simulations are stimulated by the theoretical results in Sections 4.1.1 and 4.1.2 where $K_1 = N/2 - 1$. The first case (Section 5.1) studies if the stratification on $\boldsymbol{X}_1$ in the stratified rerandomization may prohibit more efficient inference that could be achieved by rerandomization on $\boldsymbol{X}$ for $N = 16$. The second simulation (Section 5.2) considers the situation when $N = 64$ and $\beta_2 = 0$. Because of the large number of degrees of freedom in the Chi-square distribution and because x_2 does not contribute to this design, it can be seen as a 'worst case' scenario for the rerandomization design. Finally, the third simulation (Section 5.3) compares the designs in a moderately large sample size, $N = 28$, with $K_1 = K_2 = 1$ and shows how the power is affected by increasing $\mathcal{I}$, or decreasing $p_{a_0}^{\mathcal{I}}$.

We have also conducted the same MC simulations with $x_2 \sim N(0, 0.25)$, $\epsilon_i \sim N(0, 0.5)$ and with heterogeneous effects with a mean of 0.6. The results from these Monte Carlo simulations are very similar to the ones discussed below and can be obtained upon request.

5.1. MC simulation 1

The study aims at examining the power of the FRT when $\mathcal{H}$ approaches $\mathcal{N}_S$ for $N = 16$ and $K_1 = 7$. Here $\lambda_1 = \lambda_2 = 2$, $\boldsymbol{\beta}_1 = (\sqrt{\rho\zeta_1}, \ldots, \sqrt{\rho\zeta_1})'$ and $\beta_2 = \sqrt{(1-\rho)\zeta_2}$, where ζ_1 and ζ_2 are chosen such that $\boldsymbol{\beta}_1' cov(\boldsymbol{X}_1)\boldsymbol{\beta}_1 = \zeta_2^2 Var(x_2) = 0.5 \times Var(\epsilon)$. We let $\rho = R_1^2/R^2$ take the values 0, 0.5, and 1, which correspond to the binary covariates having no effect on the outcome, the binary and continuous covariates having equal effect on the outcome, and, the continuous covariate having no effect on the outcome. $\mathcal{H}$ is varied as 200, 240, and 256. Note that when $\mathcal{H} = \mathcal{N}_S = 256$, the stratified rerandomization is equal to stratification as no allocations can be excluded in the rerandomization step.

The performance under the experimental designs MR_0 and SR is considered. We denoted the Mahalanobis-based rerandomization using $\boldsymbol{X} = (\boldsymbol{X}_1, X_2)$ MR_0 as we in this simulation also consider an additional one-step rerandomization design based on only the three main covariates (i.e., the interactions among the binary covariates are excluded) which we denote MR_1. MR_1 is considered to illustrate the flexibility with rerandomization as opposed to stratification; the interactions can conveniently be included or excluded based on prior beliefs of their importance. MR_1 is a reasonable design if no a priori information about the covariates relative importance is available. That is, it can be argued that it is not reasonable to include all interactions of a set of covariates solely because they are binary. Note that the interaction terms are in fact informative when $\rho > 0$ since their coefficients are non-zero, implying that this setting does not favor MR_1 by construction. For each sample, the globally best $\mathcal{H}$ allocations according to each design are chosen.

Let $P_m(\text{SR})$, $P_m(\text{MR}_0)$ and $P_m(\text{MR}_1)$ be the estimated power in replicate m (defined in (17)) of the three designs. Fig. 3 displays the distributions, as box plots, of the relative difference in power of rerandomization compared to stratified rerandomization, defined as

$$\text{RD}_m(\text{MR}_0) = \frac{P_m(\text{MR}_0) - P_m(\text{SR})}{P_m(\text{SR})} \text{ and } \text{RD}_m(\text{MR}_1) = \frac{P_m(\text{ MR}_1) - P_m(\text{SR})}{P_m(\text{SR})}.$$

It is clear that in the settings where the binary covariates affect the outcome (panel 2 and 3), the power of MR_0 is exactly the same as SR for $\mathcal{H} = 200$, and larger when $\mathcal{H}$ comes close to $\mathcal{N}_S$. MR_1 has higher power both when $\rho = 0$ and, perhaps more surprisingly, when $\rho = 0.5$. When the continuous covariate has no effect on the outcome, MR_0 gives the same power as SR on average. Surprisingly, MR_1 perform slightly better than SR on average, also in this setting. Clearly, the information about $\boldsymbol{Y}$ sacrificed when excluding the interaction terms to lower the degrees of freedom in the rerandomization step, increases efficiency in this special case.

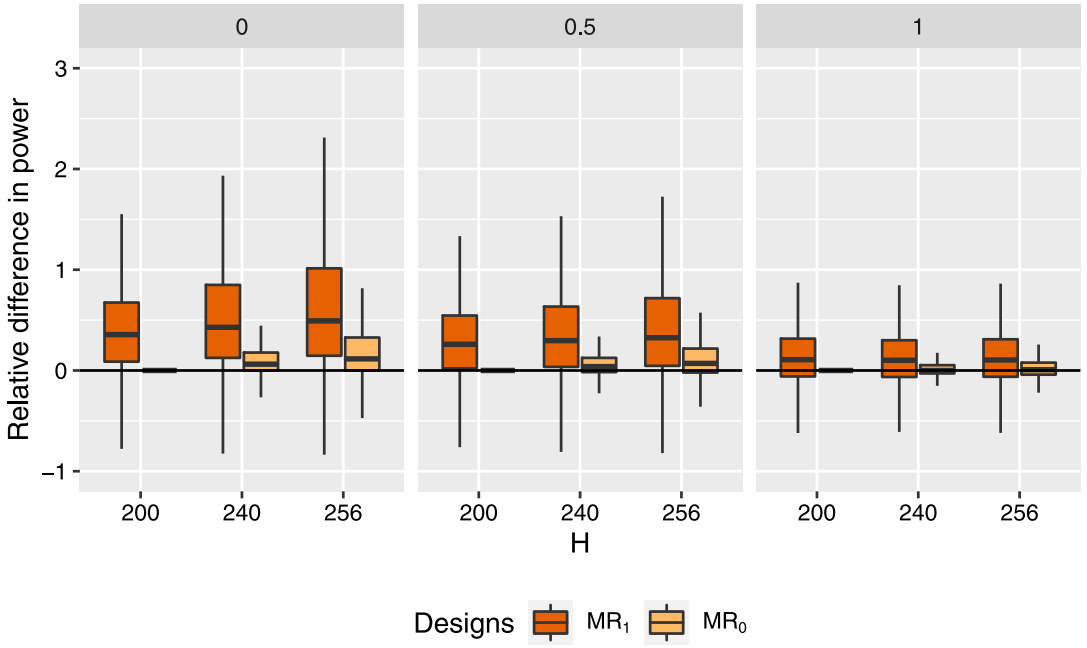

Fig. 3. The distribution, over 5000 replicates, of the relative difference in power for rerandomization as compared to stratified rerandomization with the full set of covariates and a restricted set (MR_0 and MR_1), as $\mathcal{H}$ approaches $\mathcal{N}_S = 2^8 = 256$. The panels display, from left to right, $\rho = 0$, $\rho = 0.5$, and $\rho = 1$, respectively.

This small simulation study shows that the theoretical results of Section 4.1.1 translate quite well to small samples and indicates that the theoretical results are robust to violations of the normality assumptions. If the sample size is small, the stratification reduces the number of allocations to a small set. It is then important to think about the covariates' relative importance before deciding on a design.

5.2. MC simulation 2

Here $N = 64$, $\beta_2 = 0$ and $\boldsymbol{x}_{i1}$ is a set of independent binary covariates and their interactions which imply 32 strata of size 2. λ_1 and λ_2 is chosen to obtain $var(x_2) = 0.4$ and $var(\epsilon) = 0.8$, respectively. Here $\boldsymbol{\beta}_1 = (\sqrt{\zeta_1}, \ldots, \sqrt{\zeta_1})'$ where ζ_1 is chosen such that $\boldsymbol{\beta}_1' cov(\boldsymbol{X}_1)\boldsymbol{\beta}_1 = Var(\epsilon)$. We set $\mathcal{H} = 40$, and vary $\mathcal{I} = 300, 1,000, 10,000, 20,000$, which means that $p^{\mathcal{I}}_{a_0}$ is in the range 0.133 (= 40/300) to 0.002 (= 40/20000). For CR and S, $\mathcal{H}$ allocations are randomly drawn from $\boldsymbol{W}$ and $\boldsymbol{W}^S$, respectively. For SR, $\mathcal{I}$ allocations are randomly drawn from $\boldsymbol{W}^S$ and the $\mathcal{H}$ allocations with the smallest Mahalanobis distance on x_2 within this set are chosen. For MR_0, $\mathcal{I}$ allocations are randomly drawn from $\boldsymbol{W}$ and the $\mathcal{H}$ allocations with the smallest Mahalanobis distance on $\boldsymbol{X}$ are chosen.

The maximum number of considered allocations in each replication, $\mathcal{I} = 20{,}000$, is very far from $\mathcal{N}_S = 2^{32} = 4.295 \times 10^9$. This means that there is no restriction on $\mathcal{N}^S$ in the rerandomization on x_2 in the second stage as was the case in Section 4.1.1. Instead, due to the large degrees of freedom in the Chi-square distribution in the rerandomization, and by the fact that $\beta_1 = 0$, this Monte Carlo simulation illustrates the potential problems with rerandomization in comparison to stratification and stratified rerandomization when $\boldsymbol{X}_1$ contains a large number of covariates. The degrees of freedom in the MR_0 design is 32 (31 binary, 1 continuous). This implies that, even though SR and MR_0 should give approximately equal designs asymptotically (see Eq. (11)) $p^{30,000}_{a_0} = \Pr(\chi^2(K_0) \leq a_0 | 30,000)$ is far from $\lim_{\mathcal{I}\to\mathcal{N}_A} p^{\mathcal{I}}_{a_0} = p_{a_0}$. Fig. 4 displays the distribution (box plots) of the estimated power across replications in the FRT for the four designs; $P_m(\text{C})$, $P_m(\text{S})$, $P_m(\text{SR})$ and $P_m(MR_0)$. As expected, stratification and stratified randomization achieves the full efficiency gain already with $\mathcal{I} = 300$, and the stratified randomization does not improve by rerandomizing on x_2 but is not distorted either. The rerandomization design do improves slowly with $\mathcal{I}$. However, as expected, the improvement is hardly visible across the span of $\mathcal{I}$ presented here. Table 2 displays the empirical variance of DM estimator for each design over $\mathcal{I}$, averaged over the replications. As expected, the only design for which the variance decreases with $\mathcal{I}$ is MR_0. That is, for S, and SR the full maximum variance reduction of the DM estimator, in this case $100 \times R^2 = 50\%$, is as expected achieved for all $\mathcal{I}$, whereas MR_0 only achieve 25% for $\mathcal{I} = 20{,}000$. This is in line with Fig. 2, from which we expect that with $K_0 = 32$ and $\mathcal{I} = 20{,}000$ we should have on average 60% variance reduction in $\boldsymbol{X}$, i.e., $\nu_0 = 0.6$, implying $PRIV_1 = 20\%$, i.e. $PRIV_1 = 100 \times R^2(1 - \nu_0) = 100 \times 0.5 \times 0.4$.

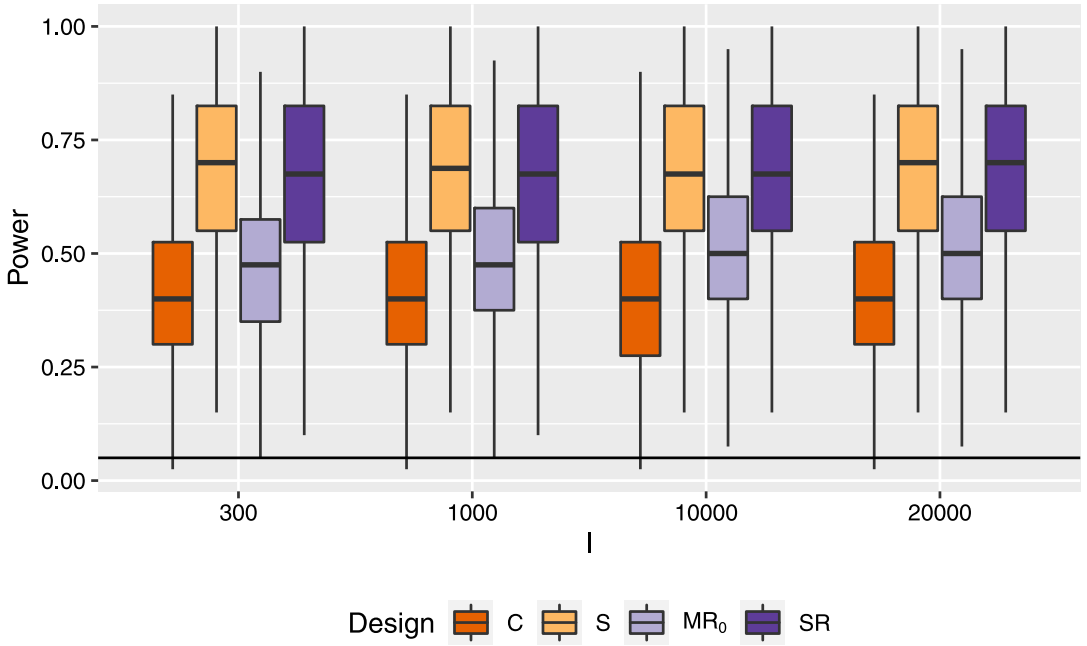

Fig. 4. The distribution, over 5000 replicates, of the power of complete randomization (C), stratification (S), rerandomization (MR_0) and stratified rerandomization (SR) as functions of number of considered allocations, $\mathcal{I}$.

Table 2
Empirical variance of the DM estimator averaged over replications for complete randomization (C), rerandomization (MR_0), stratification (S), and stratified rerandomization (SR) as a functions of number of considered allocations, $\mathcal{I}$.

$\mathcal{I}$	C	S	MR_0	SR
300	0.099	0.050	0.086	0.050
1000	0.099	0.050	0.081	0.050
10,000	0.098	0.050	0.074	0.050
20,000	0.099	0.049	0.074	0.050

5.3. MC simulation 3

Here $\beta_1 = \sqrt{\rho 2}$ and $\beta_2 = \sqrt{(1-\rho)2}$, where ρ is varied as 1/3, 1/2 and 2/3. This means that the binary covariate is half, equally and twice as important in explaining the outcome as the continuous. Furthermore, we let $\lambda_1 = 2$, $\lambda_2 = \sqrt{2}$, $N \equiv 28$, and $\mathcal{H} \equiv 40$. We vary $\mathcal{I}$ by letting $\mathcal{I} = 60, 100, 500$ and 800, which means that $p_{a_0}^{\mathcal{I}}$ varies in the range 0.67 ($= 40/60$) to 0.05 ($= 40/800$). The sampling of the $\mathcal{I}$ allocations is performed as in Section 5.2.

Fig. 5 displays the distribution (box plots) of $P_m(\mathrm{C})$, $P_m(\mathrm{S})$, $P_m(\mathrm{SR})$ and $P_m(\mathrm{MR}_0)$ across $\mathcal{I}$. As expected, the stratified design does not improve by increasing $\mathcal{I}$. All gain in efficiency from stratification is immediate since only allocations allowed under stratification are allowed. The stratified rerandomization is always better or equally good as rerandomization. This is expected as in the stratified rerandomization only allocations allowed under stratification on x_1 are allowed, and therefore it immediately starts balancing on x_2. When $\mathcal{I}$ becomes larger there is no difference between the stratified rerandomization and the rerandomization as is expected from the theoretical results derived above. The small $\mathcal{I}$ needed to obtain good power improvements in SR and MR_0 is because the number of covariates in the rerandomization is only 1 and 2, respectively.

5.4. Summary of the Monte Carlo study

The Monte Carlo simulations show that the theoretical findings apply in finite sample settings and indicate that they are robust to violations of the normality assumption of the covariate means and error term (needed to derive PRIV_t), $t = 0, 1, 2$ for the rerandomization. If the sample size is small, stratified designs can be suboptimal. If the sample size is moderately large, around 30, rerandomization without stratification run into problems when the number of strata is large. In such situations, stratified rerandomization is a good strategy for reducing the number of degrees of freedom in the Chi-square distribution of the Mahalanobis distance. This enables the Mahalanobis-based rerandomization design to benefit from informative continuous covariates. If the total number of covariates is small, say less than 5, rerandomization

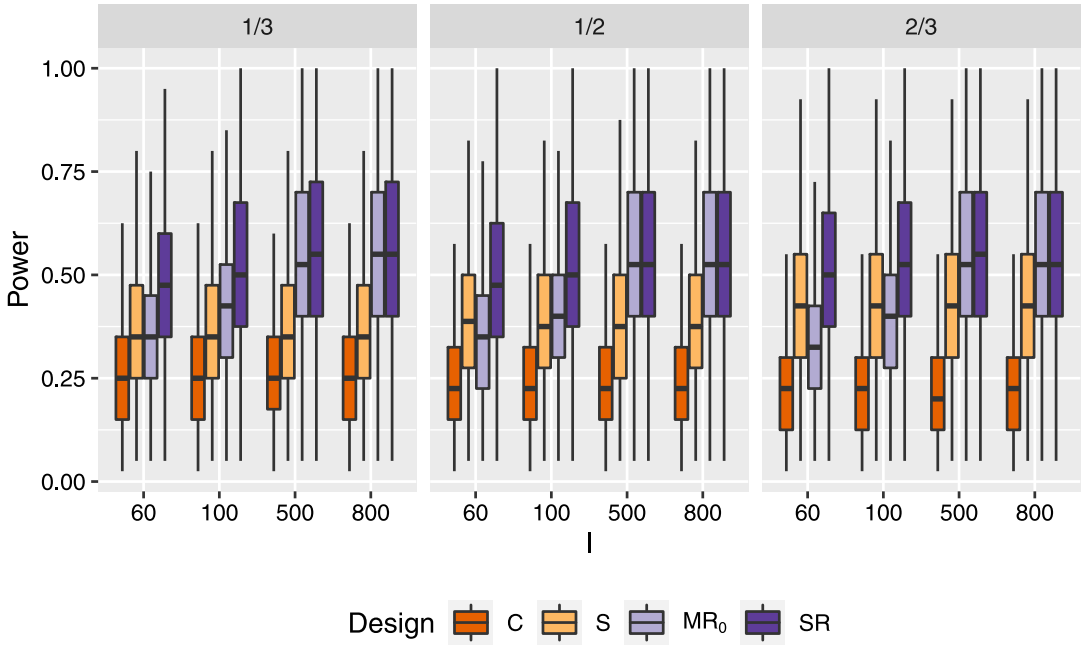

Fig. 5. The distribution, over 5000 replicates, of the power of complete randomization (C), stratification (S), rerandomization (MR_0) and stratified rerandomization (SR) as a function of the number of considered allocations $\mathcal{I}$. The panels display, from left to right, $\rho = 1/3$, $\rho = 1/2$, and $\rho = 2/3$, respectively.

and stratified rerandomization gives very similar efficiency gains in comparison to complete randomization also for small N and number of considered allocations $\mathcal{I}'s$ that are manageable for ordinary computers.

6. Empirical example

In this section we make use of an electricity consumption data set, explored in Lundgren and Schultzberg (2019). Within the field of electricity-use research, there is an increasing focus on how to change users' electricity consumption to reduce peak load as well as total load to enable further integration of weather-dependent electricity production as well as to cope with an increase in demand. Several types of interventions have been proposed and evaluated, such as financial incentives in the form of dynamic price signals (Öhrlund et al., 2019; Faruqui et al., 2017) and non-financial incentives in the form of information campaigns or energy feedback (Darby, 2006; Karlin et al., 2015). To help plan future experiments, Lundgren and Schultzberg (2019) present an exploratory study aimed at evaluating the prospects of interventions targeting the attitudes towards electricity use, and savings in particular. Due to the large natural variation in households' electricity consumption, it is crucial to employ rigorous experimental designs to achieve acceptable power, which makes this data set suitable for illustrating the designs discussed in this paper.

Electricity consumption data were collected at the monthly level for several years for 510 households of which we used the 102 with no missing data. We extracted the two last time periods, November and December for 2017 where the November measurements are being used in the experimental design. The electricity consumption in December 2017, Y_{Dec}, is the outcome. As there were no experiments in December, this electricity consumption is what is observed under the Fisher sharp null. This fact enables us to study the relative performance of the different experimental designs under the alternative and calculate the power for hypothetical treatment effects as described in Section in 5, Eqs. (16) and (17).

The November data contains the electricity consumption, Y_{Nov}, and the number of residents in each household, *Residents*. Out of the 102 households, there were 43 with 1 resident, 30 with 2 residents, 11 with 3 residents, 16 with 4 residents, 1 with 5 residents, and 1 with 6 residents.

The following six designs were considered: (i) Stratification (S) on *Residents*,[6] (ii) Stratification (S_1) on the quantiles of Y_{Nov} (Y^{Q}_{Nov}) and *Residents*,[7] (iii) Stratified rerandomization (SR) where we stratify on *Residents* and then using Mahalanobis-based rerandomization on Y_{Nov}, (iv) Rerandomization (MR_0) on Y_{Nov} and the incidence matrix implied by *Residents*, (v) Rerandomizing (MR_1) on Y_{Nov} and *Residents* and (vi), finally, Complete randomization (C) was performed as a benchmark.

[6] Since only one household each has 5 and 6 residents, respectively, only 5 strata were created merging 5 and 6 for all designs using stratification

[7] This design implies a maximum of 16 strata, in this case a few strata had zero units resulting in 13 strata.

Table 3
Expected variance reduction in the covariates under stratified rerandomization (SR) and Mahalanobis-based rerandomization (MR_0 and MR_1).

	Expected PRIV			
$p_{a_1} = p_{a_0}$ Covariate	0.15	0.015	0.0015	0.00015
Stratified rerandomization (SR)				
Residents (factor)	100	100	100	100
Y^Q_{Nov}	98.81	99.99	100	100
Rerandomization (MR_0)				
Residents (factor) and Y_{Nov}	73.45	90.75	96.48	98.62
Rerandomization (MR_1)				
Residents and Y_{Nov}	92.09	99.25	99.92	99.99

In this example, the strata are not all evenly sized which means that Corollary 3.1 does not apply. To adjust for this, the Fisher test in (i) and (ii) could be based on the stratified estimator (2). For this simulation, the analyses for the stratified designs (S, S_1, and SR) were performed with the both the stratified estimator and the DM. As expected, there are small power gains from using the stratified estimator for the stratified designs in this case as compared to the DM estimator. However, here only the DM-based results are shown to be aligned with the theoretical results in this paper. Code for reproducing the simulations, including results using the stratified estimator, is available on request.

In this special case, where the outcome under no treatment is observed, the power under hypothetical treatment effects can be studied. In a real experiment, however, only covariates are observed. However, Morgan and Rubin (2012) show that with Mahalanobis-based rerandomization the variance reduction on the observed outcome is R^2 times the variance reduction in the covariates. Since, R^2 is unknown but fixed, this gives a valid relative comparison between the design as long as $R^2 > 0$.

Table 3 displays the expected PRIV in the covariates under the different designs. It is clear that if only a small set of all allocations is considered, there are large differences in the variance reduction in the covariates across the designs. For $\mathcal{I} = 10^5$ ($p^{\mathcal{I}}_{a_1} = 0.15$) the variance reduction in the SR design (iii) is 100% and 98.8% for Residents and Y_{Nov}, respectively. As the same variance reduction is obtained for all covariates by definition with the Mahalanobis distance measure, the corresponding variance reduction for MR_0 and MR_1 is 73.45% and 92.09%, respectively. When the number of allocations $\mathcal{I} = 10^8$ (i.e. $p^{\mathcal{I}}_{a_1} = p^{\mathcal{I}}_{a_0} = 0.00015$) we get almost 100% variance reductions on all covariates for these three designs.

Returning to simulating the power under hypothetical effects, we let

$$Y_i(0) = Y_{i,Dec}$$
$$Y_i(1) = Y_i(0) - \tau,$$

where τ is varied as 0, 10, 20, and 30 (kWh), which correspond almost exactly to 0, 0.1, 0.2, and 0.3 standard deviation of $\boldsymbol{Y}_{Dec}$ ($s_{\boldsymbol{Y}_{Dec}} = 102.2$). The negative sign of the effect is motivated by the intervention aiming at decreasing consumption.

We follow the procedure in Section 5 but let $\mathcal{H} = 15{,}000$ ($= 2/r$) which implies that $r = 0.00013$. For the Mahalanobis distances rerandomization we randomize among the $\mathcal{H}$ allocations with smallest Mahalanobis distances in a random set of sizes $\mathcal{I} = 10^5, 10^6, 10^7$ and 10^8. This means that $p^{\mathcal{I}}_{a_0}$ and $p^{\mathcal{I}}_{a_1}$ vary in the range 0.15 to 0.00015. For the C, S and S_1 designs, 15,000 allocations were randomly drawn from $\boldsymbol{W}$. For the SR design the randomization is conducted in the $\mathcal{H}$ allocations with the smallest Mahalanobis in the random set from $\boldsymbol{W}^S$ of sizes $\mathcal{I} = 10^5, 10^6, 10^7$ and 10^8.[8]

Fig. 6 displays the power of the five designs and complete randomization for increasing number of considered allocations, $\mathcal{I}$. For $I = 10^5$ ($p^{\mathcal{I}}_{a_0} = p^{\mathcal{I}}_{a_1} = 0.15$) there are large differences between the different rerandomization designs and complete randomization. Among the designs using rerandomization, the price of the degrees of freedom in the rerandomization designs, discussed in Section 4.1.2, is clearly seen as MR_0 has the lowest power, MR_1 is in the middle, and SR has the highest power. When $\mathcal{I} = 10^8$ ($p^{\mathcal{I}}_{a_0} = 0.00015$) these differences are negligible. Stratification on Y^Q_{Nov} and *Residents* gives substantial power improvements as compared to CR. For $\mathcal{I} = 10^5$, S_1 has higher power than MR_0, however, as S_1 does not improve with $\mathcal{I}$, this does not hold when $\mathcal{I}$ increases. The difference between SR and S_1 clearly illustrates the (unnecessary) information loss associated with discretizing Y_{Nov}. S also gives higher power than C, but as expected, the importance of balancing the pre-treatment outcome is far more rewarding than perfect balance on number of residents.

[8] Note that an alternative is to conduct a Monte Carlo approximations from the set $\boldsymbol{W}$. The number of Monte Carlo draws needs to be large in order for the FRT to have the right level. In a single analysis this is not a problem, however, in a Monte Carlo simulation this procedure would be very time consuming.

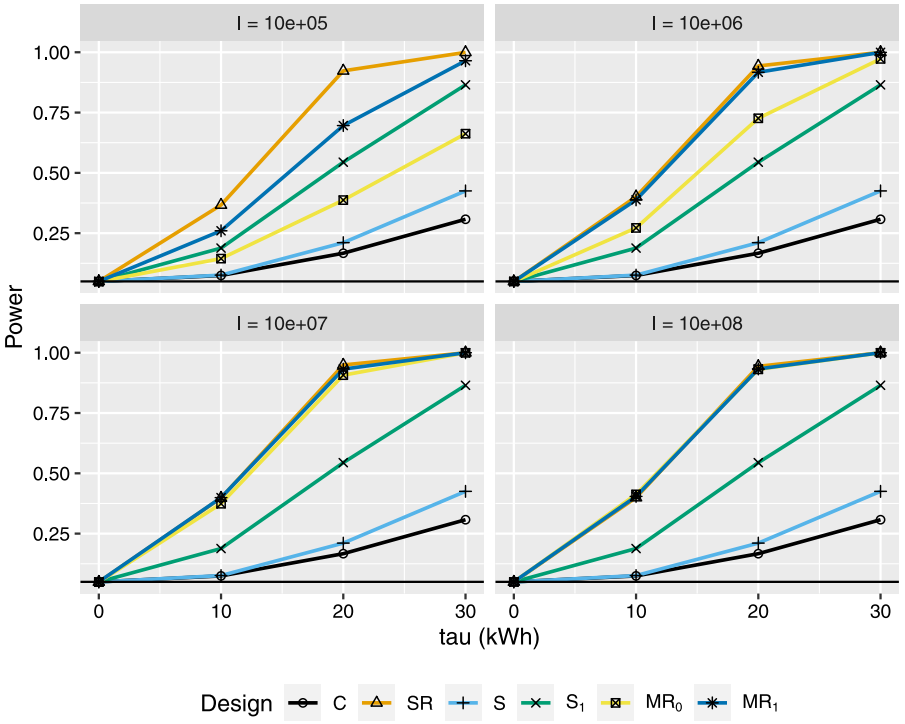

Fig. 6. Power of FRT's under hypothetical homogeneous treatment effects for the five designs (Stratification (S) on *Residents*, Stratification (S_1) on Y^Q_{Nov} and *Residents*, Stratified rerandomization (SR) where we stratify on *Residents* and then using Mahalanobis-based rerandomization on Y_{Nov}, Mahalanobis-based rerandomization on Y_{Nov} and the incidence matrix implied by *Residents* (MR_0), Mahalanobis-based rerandomizing on Y_{Nov} and *Residents* (MR_1) and complete randomization (C).

7. Discussion

Stratification, or blocked randomization, is the most common design used to improve balance in experiments. An alternative, or complement, to blocking that has received attention lately is to utilize modern computational capabilities in finding allocations with balance in observed covariates (see e.g., Morgan and Rubin, 2012, 2015; Bertsimas et al., 2015; Kallus, 2018; Lauretto et al., 2017; Krieger et al., 2019; Kapelner et al., 2020; Johansson and Schultzberg, 2020).

Several scholars, including R A. Fisher (via Cochran and Rubin), and more recently Morgan and Rubin (2012, 2015), recommend rerandomization as a complement to stratification. Athey and Imbens (2016) on the other hand seem to view rerandomization as an alternative rather than a complement, and do not recommend rerandomization.

It is, however, not obvious how or when to combine these strategies. The paper has investigated the properties and limitations of *stratification*, *rerandomization*, and the combinations of stratification and rerandomization, denoted *stratified rerandomization*, with the aim of clarifying their pros and cons in practice. The comparison is focused on the efficiency of the differences-in-mean estimator under homogeneous treatment effects and inference to the experiment, that is for a fixed N. We show that when all strata are of even sample size the stratified design give the same design as with Mahalanobis-based rerandomization with a criterion set to zero. By using the results in Morgan and Rubin (2015), who uses the Mahalanobis distance between the means of the covariates of 'treated' and 'controls' in the rerandomization, this enables us to study computational aspects and the relative efficiency of *stratified rerandomization* and *rerandomization*.

The main conclusion is that there are three aspects to consider when choosing between these two designs: (i) the number of available binary (on which it is easy to stratify) and continuous covariates, (ii) the relative importance on the outcome of these two types of covariates , and (iii) the number of allocations remaining after stratification. If the number of allocations remaining after a stratification is large, say 10,000 or larger, it is a good idea to use *stratified rerandomization*. If the number of allocations remaining after stratification is small, it is possible that rerandomization on all covariates at once, or only the continuous covariates, is more efficient. If the sample size is moderately large, say 20 or larger, and the total number of covariates is small, say 5 or less, the efficiency gain of rerandomization compared to stratified rerandomization is negligible.

The results are of direct interest for researchers that are restricted to conduct small experiment. The results are however also of interest for researchers conducting large experiment. With large experiment it is always advisable to stratify and then to further increase efficiency by rerandomizing on essential covariates within each stratum. The Fisher exact test can then be based on the stratified estimator (2).

How asymptotic inference should be conducted is something that we leave to future research. However, it seems quite straight forward by extending the results in Li et al. (2018) or Li and Ding (2020). With the Li et al. (2018) approach the inference would be based on the stratified estimator and with the Li and Ding (2020) approach inference would be based on the regression adjustment estimator suggested in Imbens and Rubin (2015, p. 206).

References

Athey, Susan, Imbens, Guido, 2016. The state of applied econometrics - causality and policy evaluation. 31, (2), (ISSN: 0895-3309) pp. 3–32. http://dx.doi.org/10.1257/jep.31.2.3, URL http://arxiv.org/abs/1607.00699.

Bertsimas, Dimitris, Johnson, Mac, Kallus, Nathan, 2015. The power of optimization over randomization in designing experiments involving small samples. Oper. Res. 63 (4), 868–876. http://dx.doi.org/10.1287/opre.2015.1361.

Darby, Sarah, 2006. The effectiveness of feedback on energy consumption. In: A Review for DEFRA of the Literature on Metering, Billing and Direct Displays. Vol. 486.

Faruqui, Ahmad, Sergici, Sanem, Warner, Cody, 2017. Arcturus 2.0: A meta-analysis of time-varying rates for electricity. Electr. J. (ISSN: 10406190) 30 (10), 64–72. http://dx.doi.org/10.1016/j.tej.2017.11.003, URL http://linkinghub.elsevier.com/retrieve/pii/S1040619017302750.

Imbens, Guido W., Rubin, Donald B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, ISBN: 9780521885881.

Johansson, Per, Schultzberg, Mårten, 2020. Rerandomization strategies for balancing covariates using pre-experimental longitudinal data. J. Comput. Graph. Statist. 29 (4), 798–813. http://dx.doi.org/10.1080/10618600.2020.1753531.

Kallus, Nathan, 2018. Optimal a priori balance in the design of controlled experiments. J. R. Stat. Soc. Ser. B Stat. Methodol. (ISSN: 14679868) 80 (1), 85–112. http://dx.doi.org/10.1111/rssb.12240, http://arxiv.org/abs/1312.0531,

Kapelner, Adam, Krieger, Abba M., Sklar, Michael, Shalit, Uri, Azriel, David, 2020. Harmonizing optimized designs with classic randomization in experiments. Amer. Statist. 1–12. http://dx.doi.org/10.1080/00031305.2020.1717619.

Karlin, Beth, Zinger, Joanne F., Ford, Rebecca, 2015. The effects of feedback on energy conservation: A meta-analysis. Psychol. Bull. (ISSN: 00332909) 141 (6), 1205–1227. http://dx.doi.org/10.1037/a0039650.

Krieger, A.M., Azriel, D., Kapelner, A., 2019. Nearly random designs with greatly improved balance. Biometrika (ISSN: 0006-3444) 106 (3), 695–701. http://dx.doi.org/10.1093/biomet/asz026.

Lauretto, Marcelo S., Stern, Rafael B., Morgan, Kari L., Clark, Margaret H., Stern, Julio M., 2017. Haphazard intentional allocation and rerandomization to improve covariate balance in experiments. AIP Conf. Proc. (ISSN: 15517616) 1853 (June), http://dx.doi.org/10.1063/1.4985356,

Li, Xinran, Ding, Peng, 2020. Rerandomization and regression adjustment. J. R. Stat. Soc. Ser. B Stat. Methodol. 82 (1), 241–268. http://dx.doi.org/10.1111/rssb.12353, http://arxiv.org/abs/https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12353.

Li, Xinran, Ding, Peng, Rubin, Donald B., 2018. Asymptotic theory of rerandomization in treatment-control experiments. Proc. Natl. Acad. Sci. 115 (37), 9157–9162.

Lundgren, Berndt, Schultzberg, Mårten, 2019. Does Energy-Effective Behavior Matter for Energy Conservation?. KTH Royal Institute of Technology.

Morgan, Kari Lock, Rubin, Donald B., 2012. Rerandomization to improve covariate balance in experiments. Ann. Statist. (ISSN: 00905364) 40 (2), 1263–1282. http://dx.doi.org/10.1214/12-AOS1008, http://arxiv.org/abs/arXiv:1207.5625v1.

Morgan, Kari Lock, Rubin, Donald B., 2015. Rerandomization to balance tiers of covariates. J. Amer. Statist. Assoc. (ISSN: 1537274X) 110 (512), 1412–1421. http://dx.doi.org/10.1080/01621459.2015.1079528,

Öhrlund, Isak, Schultzberg, Mårten, Bartusch, Cajsa, 2019. Identifying and estimating the effects of a mandatory billing demand charge. Appl. Energy (ISSN: 03062619) 237, http://dx.doi.org/10.1016/j.apenergy.2019.01.028.

Rubin, Donald B., 1980. Randomization analysis of experimental data: The Fisher randomization test comment. J. Amer. Statist. Assoc. (ISSN: 01621459) 75 (371), 591–593. http://dx.doi.org/10.2307/2287653.