Received: 10 May 2022 | Revised: 9 November 2022 | Accepted: 16 January 2023

DOI: 10.1002/sim.9672

RESEARCH ARTICLE

Statistics in Medicine WILEY

Simulated annealing for balancing covariates

Alessandro Baldi Antognini | **Marco Novelli** | **Maroussa Zagoraiou**

Department of Statistics, University of Bologna, Bologna, Italy

Correspondence
Marco Novelli, Department of Statistics, University of Bologna, Via Belle Arti 41, 40126, Bologna, Italy.
Email: m.novelli@unibo.it

Covariate balance is one of the fundamental issues in designing experiments for treatment comparisons, especially in randomized clinical trials. In this article, we introduce a new class of covariate-adaptive procedures based on the Simulated Annealing algorithm aimed at balancing the allocations of two competing treatments across a set of pre-specified covariates. Due to the nature of the simulated annealing, these designs are intrinsically randomized, thus completely unpredictable, and very flexible: they can manage both quantitative and qualitative factors and be implemented in a static version as well as sequentially. The properties of the suggested proposal are described, showing a significant improvement in terms of covariate balance and inferential accuracy with respect to all the other procedures proposed in the literature. An illustrative example based on real data is also discussed.

KEYWORDS

covariate-adaptive procedures, loss of information, mahalanobis distance, rerandomization, treatment comparisons

1 | INTRODUCTION

In comparative experiments, balancing the covariates across experimental groups is a crucial requirement to ensure the credibility of the trial results and to guarantee optimal inference about the treatment effects.[1-3] In this regard, several procedures have been suggested, with the aim of creating comparable treatment groups with respect to the selected prognostic factors. One of the oldest approach is the so-called minimization method for qualitative factors, tracing back to the work of Taves[4] and Pocock and Simon[5] and later generalized by Hu and Hu,[6] intended to minimize a weighted sum of the marginal imbalances of allocations for all covariates. Another well-known approach instead exploits stratification by making use of a separate randomization procedure within each stratum in order to improve balance.[7] Both methods are suitable only for categorical variables, while continuous covariates are either ignored or discretized: the former solution leads to potentially serious power loss,[8] while the latter one, despite being widespread in clinical research, may strongly damage the inferential precision since the nature of the variables changes due to the subjective choices of the thresholds.[9-11] Moreover, as the number of covariates increases, the stratification approach may rapidly become impractical especially for small sample sizes, while minimization method may cause practical problems in real life applications due to the increasing required complexity.[3,9]

Adopting an optimal design perspective, Atkinson[12] and Smith[13,14] suggested covariate adaptive procedures directed at minimizing the variance of the estimated treatment effect, under the classical linear homoschedastic model setup. Albeit these rules are compatible with both categorical and continuous covariates, in general their performances are

strongly related to both the correctness of the model specification[15] and, as also our results show, its complexity. Finally, an alternative approach introduced by Ma and Hu[16] is aimed at minimizing a weighted average of the estimated distributional imbalances obtained by means of kernel density estimations. For a complete overview see Rosenberger and Sverdlov,[1] Rosenberger and Lachin[17] and Baldi Antognini and Giovagnoli.[18]

In the causal inference framework when all patients' covariates are available before the experiment starts, Morgan and Rubin[2] introduced the so-called rerandomization (RR) approach for balancing quantitative covariates across two experimental groups. Under this procedure, adopting complete randomization units are repeatedly allocated to the treatments until a prefixed covariate balance criterion is satisfied. In particular, the authors considered as the imbalance measure the Mahalanobis distance between the sample means across the two groups and proposed to stop the process and accept the treatment allocation when the distance falls below some prefixed constant, which specifies the maximum amount of tolerated imbalance. In order to deal with sequential enrollment designs, RR was later generalized in a group sequential way.[19] As the number of the considered covariates increases, the computational burden required to obtain an acceptable configuration increases too: this may lead to a compromise between computational feasibility and covariate balance. Moreover, this procedure should be applied only when all or at least some of the variables are quantitative.[19]

In the last decade, thanks to the recent advances in the biomarkers-based personalized medicine, it has become increasingly common to include several covariates and their interactions in the analysis.[20-25] However, with the exception of Atkinson's procedure that induces a lower order balance,[7] an efficient Covariate-Adaptive (CA) procedure able to deal with mixed covariates profile with potentially complex interaction structure is still missing.

In this paper, we propose a new class of designs based on the Simulated Annealing (SA) algorithm, which is aimed at balancing the allocations among a set of pre-specified covariates. Originally suggested in the context of statistical mechanics,[26] SA is a stochastic local search algorithm which has been vastly used to approximate global optimization solutions for large search spaces.[27-29] It comes from the physical process of the annealing of metals by gradual cooling: at high temperatures, the particles are rather free to move, leaving the structure subject to substantial changes, while as the temperature gradually decreases, the probability that a particle will move decreases too, until the system reaches a steady state. In a nutshell, the algorithm starts from some initial point and then it iteratively explores its neighborhood; better solutions will be always accepted, while worsening ones are retained probabilistically, depending on both the amount of deterioration and a parameter called temperature, that governs the evolution of the procedure. Large temperatures allow the algorithm to search for new potential optimal solutions in a wider area, so the probability of identifying the global minimum tends to grow, while low values, by inducing a smaller search area, may increase the risk of being trapped in local minima. Due to the stochastic nature of the SA algorithms, upward moves (namely worse solutions) can be occasionally accepted: this is done in the hope that such choices will allow the algorithm to escape from local minima, in order to find the global optimum. Markov chains are the underlined probabilistic models that govern the behavior of the SA algorithm, which converges in probability to the global optimum as the number of iterations grows (under widely satisfied conditions).[29-31]

In this article, the SA algorithm is exploited to control covariate imbalance. In particular, we introduce a new class of CA procedures called the simulated annealing designs (SADe) which:

- can deal with continuous and/or categorical variables,
- allow the adoption of any specific measure of covariate imbalance,
- can be applied to both fixed (ie, nonsequential) experiments, where all the covariate information is available before the trial begins, and sequential ones in which statistical units enter the trial progressively, also allowing for a group sequential version,
- turn out to be remarkably effective in the case of small sample sizes and a large number of covariates, a critical set-up in which enforcing covariate balance is particularly important.

Moreover, due to the stochastic nature of the SA algorithm, SADe are intrinsically randomized and completely unpredictable, thus avoiding any possible selection bias. An extensive simulation study is performed to show the excellent performances of our proposal: the finite sample properties of the SADe are compared with those of other well-known CA procedures, by taking into account (i) stratified randomization methods, minimization procedures (and their generalization[6]) for qualitative covariates, and (ii) RR (in its fixed and group sequential versions), Atkinson's optimal design[12] and kernel density-based procedure[16] when some covariates are quantitative, also considering the completely randomized design as a benchmark. Starting from some preliminaries in Section 2, Section 3 introduces the SADe and Section 4 deals with the finite sample comparisons between CA rules, taking also into account a real case study; Section 5 shows

the impact of the covariate imbalance on the inferential accuracy and the final section discusses some guidelines for the practical application of the suggested proposal.

2 | NOTATION, IMBALANCE MEASURES AND INFERENTIAL PRECISION

We will consider experiments where two treatments, say A and B, are compared. Suppose that n assignments have been made to statistical units with independent and identically distributed (i.i.d.) covariates $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ belonging to a given covariate distribution $\mathcal{G}(\boldsymbol{x})$. Each $\boldsymbol{x}_i$ is a q-dimensional random variable representing the set of covariates (qualitative and/or quantitative), for which balance between treatment groups is desired, and are assumed to be measurable before the assignment. We denote by Y_i the experimental outcome of the ith subject, $\boldsymbol{x}_i = (x_{i1}, \dots, x_{iq})^t$ is his/her covariate profile, while δ_i represents the corresponding allocation, with $\delta_i = 1$ if the ith subject is assigned to A and 0 otherwise. From now on we set $\boldsymbol{Y}_n = (Y_1, \dots, Y_n)^t$ and $\boldsymbol{\delta}_n = (\delta_1, \dots, \delta_n)^t$, while we denote by $\mathbf{1}_n$ and $\boldsymbol{I}_n$ the n-dimensional vector of ones and the identity matrix, respectively. Let $\boldsymbol{X}_n$ be the $(n \times p)-$dimensional matrix where the chosen covariates are measured on the experimental units (usually $p = q$, but $\boldsymbol{X}_n$ may also includes power transformations and interactions so, in general, $p \geq q$) and let $\bar{\boldsymbol{x}}_{An} = \boldsymbol{X}_n^t\boldsymbol{\delta}_n/\mathbf{1}_n^t\boldsymbol{\delta}_n$ and $\bar{\boldsymbol{x}}_{Bn} = \boldsymbol{X}_n^t(\mathbf{1}_n - \boldsymbol{\delta}_n)/(n - \mathbf{1}_n^t\boldsymbol{\delta}_n)$ be the vectors collecting the sample means of the two groups.

In the framework of causal inference for quantitative (normal) covariates, Morgan and Rubin[2] proposed the RR method by assuming as a measure of covariate imbalance the Mahalanobis distance between $\bar{\boldsymbol{x}}_{An}$ and $\bar{\boldsymbol{x}}_{Bn}$, namely

$$M_n = (\bar{\boldsymbol{x}}_{An} - \bar{\boldsymbol{x}}_{Bn})^t \mathrm{var}(\bar{\boldsymbol{x}}_{An} - \bar{\boldsymbol{x}}_{Bn})^{-1}(\bar{\boldsymbol{x}}_{An} - \bar{\boldsymbol{x}}_{Bn}) = n\pi_n(1 - \pi_n)(\bar{\boldsymbol{x}}_{An} - \bar{\boldsymbol{x}}_{Bn})^t \mathrm{var}(\boldsymbol{x})^{-1}(\bar{\boldsymbol{x}}_{An} - \bar{\boldsymbol{x}}_{Bn}),$$

where $\mathrm{var}(\boldsymbol{x})$ represents the sample covariance matrix of the covariates and $\pi_n = n^{-1}\boldsymbol{\delta}_n^t\mathbf{1}_n$ is the percentage of allocations to A.

In the context of model-based inference, inspired by the linear homoscedastic model

$$E(\boldsymbol{Y}_n) = \boldsymbol{\delta}_n\theta_A + (\mathbf{1}_n - \boldsymbol{\delta}_n)\,\theta_B + \boldsymbol{X}_n\boldsymbol{\beta}, \qquad \mathrm{var}(\boldsymbol{Y}_n) = \sigma^2\boldsymbol{I}_n, \tag{1}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of covariate effects (considered as nuisance parameters) and σ^2 is the common variance of the two arms, an alternative and widely used measure of covariate imbalance is the so-called loss of information,[12]

$$\ell_n = n^{-1}\boldsymbol{b}_n^t(n^{-1}\mathbb{F}_n^t\mathbb{F}_n)^{-1}\boldsymbol{b}_n, \tag{2}$$

where $\mathbb{F}_n = [\mathbf{1}_n; \boldsymbol{X}_n]$, $\boldsymbol{b}_n^t = (D_n; (2\boldsymbol{\delta}_n - \mathbf{1}_n)^t\boldsymbol{X}_n)$ is the so-called *imbalance vector* and $D_n = 2\boldsymbol{\delta}_n^t\mathbf{1}_n - n$ is the difference between the allocations in the two groups. Basically, ℓ_n represents the loss of estimation precision induced by the covariate imbalance after n assignments, while the corresponding loss of estimation efficiency is ℓ_n/n. Under this framework the inferential goal typically consists in estimating the treatment effects, $(\theta_A; \theta_B)$, or the treatment difference, $\theta_A - \theta_B$, as precisely as possible and, by taking into account the well-known A-, D- and D_A-optimality, those criteria depend on the design only through the loss.[7,18] Indeed, after n assignments, let $\hat{\boldsymbol{\theta}}_n = \left(\hat{\theta}_{An}; \hat{\theta}_{Bn}\right)^t$ be the least square estimator of the treatment effects $\boldsymbol{\theta}$, then

$$\mathrm{tr}\left[\,\mathrm{var}\left(\hat{\boldsymbol{\theta}}_n\right)\right] = \frac{\sigma^2}{n}\left(1 - \frac{\ell_n}{n}\right)^{-1}, \quad \det\left[\mathrm{var}\left(\hat{\boldsymbol{\theta}}_n\right)\right] = \frac{4\sigma^4\left(1 - \frac{\ell_n}{n}\right)^{-1}}{n^2\left[1 - n^{-1}\bar{\mathbf{f}}_n^t\boldsymbol{X}_n^t\boldsymbol{X}_n\bar{\mathbf{f}}_n\right]} \quad \text{and} \quad \mathrm{var}(\hat{\theta}_{An} - \hat{\theta}_{Bn}) = \frac{4\sigma^2}{n}\left(1 - \frac{\ell_n}{n}\right)^{-1},$$

where $\bar{\mathbf{f}}_n = n^{-1}\boldsymbol{X}_n^t\mathbf{1}_n$ denotes the vector of the sample means for all the observations. Clearly, for every sample size n the estimation efficiency is maximized when $\ell_n = 0$; the same conclusion holds when the inferential interest lies in testing the null hypothesis $H_0 : \theta_A = \theta_B$ versus $H_1 : \theta_A \neq \theta_B$ through the classical Wald statistic $W_n = (\hat{\theta}_{An} - \hat{\theta}_{Bn})^2/\mathrm{var}(\hat{\theta}_{An} - \hat{\theta}_{Bn})$.

3 | THE SIMULATED ANNEALING DESIGN

In this Section we describe the new covariate balanced procedure based on the SA algorithm, the Simulated Annealing Design (SADe). For the sake of clarity, we start with its fixed (ie, nonsequential) version by assuming that all the patients'

covariates are available before the trial starts; then, we illustrate the SADe in its sequential version, namely by assuming that the experimental units are enrolled sequentially in groups.

3.1 | The fixed (nonsequential) scenario

For an experiment with n statistical units, let $\psi_n = \psi(\boldsymbol{\delta}_n|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) : \{0;1\}^n \to \mathbb{R}^+ \bigcup \{0\}$ be the chosen imbalance measure to be minimized (eg, the loss ℓ_n, the Mahalanobis distance M_n or other user-selected measures), so that the problem consists in finding the assignment vector $\boldsymbol{\delta}_n^* = (\delta_1^*, \dots, \delta_n^*)^t \in \{0;1\}^n$ minimizing ψ, namely such that $\psi(\boldsymbol{\delta}_n^*|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \leq \psi(\boldsymbol{\delta}_n|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ for every $\boldsymbol{\delta}_n$.

Starting from an initial allocation $\boldsymbol{\delta}_I$, which is randomly chosen in the lattice of order 2^n, the algorithm individuates the neighbor allocations on the basis of the chosen topology (eg, allowing only one or more components to vary). Then, the SA procedure randomly chooses one of them, say $\boldsymbol{\delta}_N$, and the algorithm evolves by comparing the new candidate $\boldsymbol{\delta}_N$ with $\boldsymbol{\delta}_I$ in terms of ψ, namely by evaluating the change $\Delta = \psi(\boldsymbol{\delta}_N|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) - \psi(\boldsymbol{\delta}_I|\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$. Thus, if $\Delta \leq 0$ then $\boldsymbol{\delta}_N$ is automatically accepted as the new current solution. If $\Delta > 0$ instead, the candidate $\boldsymbol{\delta}_N$ is accepted with probability $\exp\{-\Delta/T_0\}$; otherwise the algorithm stays in $\boldsymbol{\delta}_I$ and a new neighbor should be selected. In such a way the SA algorithm evolves until the prefixed number of iterations is reached. Here, $T_0 > 0$ is the initial temperature and it is customary to assume a decreasing sequence of temperatures as the iterations progress (ie, a cooling scheme $T_0 > T_1 > T_2 > \dots$), also allowing multiple iterations at each temperature level.[31] Generally, in practical applications no prior information about the optimal annealing scheme for the considered problem is available, so that many authors suggest to rely on the standard geometrical scheme, $T_{n+1} = rT_n$ where $r \in (0.8; 0.99)$ is called the temperature decay, since it has the advantage of being robust and ensures convergence toward an approximate solution.[29,32,33] Another important aspect regards the choice of the neighborhoods, namely the topology enforced by the way in which the algorithm searches for possible candidates in the neighbors of the current solution. Inspired by the work of Mladenović and Hansen,[34] Bouffard and Ferland[35] and Palubeckis,[36] where the advantages of adopting a variable neighborhood search strategy are shown, in this paper we adopt a neighborhood search scheme that resembles the cooling one. More specifically, let the ith neighborhood of the current solution $\boldsymbol{\delta}_C$ be the set of all the neighbor allocations generated by varying i components of $\boldsymbol{\delta}_C$; then as the iterations progress and temperature decreases, the algorithm makes use of a set of neighborhoods in which the number of varying entries gradually reduces. This enables for a wide search at the beginning and narrows the set of possible candidates at the end of the computation when the current solution should be close to the optimal one.

Remark 1. One of the point of strength of SADe is its flexibility, indeed there are no restrictions in the choice of the imbalance measure ψ_n, the loss or the Mahalanobis distance being just two well-known examples; this allows the experimenter to suitable choose a specific measure in accordance with the trial objectives.

3.2 | SADe for sequentially minimizing the covariate imbalance

Let us now consider the general case in which n experimental units are enrolled sequentially and, after recording their covariate patterns, they have to be assigned to A or B. In particular, assume that subjects enter the trial in groups of $m \geq 1$ statistical units; so, by denoting with $k = n/m$ the prefixed total number of groups, the ith one includes subjects from $(m(i-1)+1)$th to (mi)th, for $i = 1, \dots, k$ (note that, the groups can also have different sizes).

The sequential SADe is a procedure for minimizing the covariate imbalance that works at each step in three directions: (i) given the information accrued so far, the covariate distribution $\mathcal{G}(\boldsymbol{x})$ is estimated by $\hat{\mathcal{G}}(\boldsymbol{x})$, (ii) $\hat{\mathcal{G}}(\boldsymbol{x})$ is then employed to randomly sample the covariate pattern of the remaining (future) patients for deriving the 'predicted' imbalance measure $\tilde{\psi}$ and (iii) SA is applied to identify the treatment allocation sequence minimizing $\tilde{\psi}$, from which the assignments are extracted.

The rationale behind this procedure is that, by estimating $\mathcal{G}(\boldsymbol{x})$ and randomly generating the profiles of the remaining patients to derive the corresponding 'predicted' imbalance measure $\tilde{\psi}$, SADe avoids the characteristic behavior of CA rules. Indeed, the latter tend to make the assignment as a 'local' optimal choice conditionally to the available information accrued up to that step, which could be very far from the global optimum due to the absence of information about the covariates of the future units.

Formally, SADe can be sketched as follows:

(1) for $i = 1$: as the first group enters the trial, use $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ to derive $\hat{\mathcal{G}}_1(\boldsymbol{x})$, namely the estimated covariate distribution for the first m patients, and randomly generate the remaining $n - m$ profiles $\tilde{\boldsymbol{x}}_{m+1}, \ldots, \tilde{\boldsymbol{x}}_n$ accordingly; then, apply the SA algorithm to the predicted imbalance measure $\tilde{\psi}(\delta_1, \ldots, \delta_n \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m, \tilde{\boldsymbol{x}}_{m+1}, \ldots, \tilde{\boldsymbol{x}}_n)$ to derive the optimal assignments $\delta_1^*, \ldots, \delta_n^*$ and allocate the first group according to $\delta_1^*, \ldots, \delta_m^*$;
(2) for $i = 2, \ldots, k-1$: when the ith group is available, evaluate $\hat{\mathcal{G}}_i(\boldsymbol{x})$ by using $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{mi}$ and generate the remaining $n - mi$ profiles $\tilde{\boldsymbol{x}}_{mi+1}, \ldots, \tilde{\boldsymbol{x}}_n$ according to it, to derive

$$\tilde{\psi}(\delta_{m(i-1)+1}, \ldots, \delta_n \mid \boldsymbol{\delta}_{m(i-1)}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{mi}, \tilde{\boldsymbol{x}}_{mi+1}, \ldots, \tilde{\boldsymbol{x}}_n) \tag{3}$$

(namely $\tilde{\psi}$ corresponds to the chosen covariate imbalance measure where $\boldsymbol{\delta}_{m(i-1)}$ are the assignments actually made, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{mi}$ are the observed covariate patterns, while $\delta_{m(i-1)+1}, \ldots, \delta_n$ are the future allocations to be determined); then, apply the SA to (3) to derive $\delta^*_{m(i-1)+1}, \ldots, \delta^*_n$ and allocate the ith group of subjects according to $\delta^*_{m(i-1)+1}, \ldots, \delta^*_{mi}$;
(3) for $i = k$: apply the SA algorithm to $\psi(\delta_{n-m+1}, \ldots, \delta_n \mid \boldsymbol{\delta}_{n-m}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ to derive $\delta^*_{n-m+1}, \ldots, \delta^*_n$ and allocate the last group accordingly.

Clearly, the fully sequential version of the SADe can be easily derived by letting $m = 1$: in this case, a (small) starting sample—assigned via restricted randomization—is needed to obtain nontrivial estimates of the covariate distribution $\mathcal{G}$. Analogously, the fixed scenario previously discussed corresponds to the special case $m = n$ (namely, $k = 1$). It should be noted that, unlike the fixed case, in the sequential scenario the procedure does not need the full set of covariate profiles of the patients in advance, rather this information will be progressively available as the trial unfolds. From now on, SADe(1) and SADe(m) denote the fully sequential and the group sequential versions of the Simulated Annealing Design, respectively, while SADe* simply denotes its fixed version.

Remark 2. The performances of the Simulated Annealing Design in the sequential version (ie, for $k > 1$) clearly depend on the quality of the estimation of the covariate distribution $\mathcal{G}$. In general no specific assumption about the dependence structure of the covariates is needed. Indeed, if all the chosen covariates are categorical, we can linearize the Cartesian product of their supports into a vector of strata in a multinomial framework. For example, suppose that the jth covariate has l_j levels ($j = 1, \ldots, p$), so that $S = \prod_{j=1}^{p} l_j$ is the total number of strata; let p_s be the probability to observe the sth stratum ($s = 1, \ldots, S$), then it will be estimated at each step i by the current proportion $\hat{p}_{si}$ of units observed in this stratum. For quantitative covariates, a non-parametric multivariate kernel estimation could be employed, especially if no a-priori information about the joint distribution $\mathcal{G}$ is available. The mixed scenario, namely with both qualitative and quantitative covariates, can be derived by combining the two previous estimation approaches. As is well known, when the number of the considered quantitative covariates grows the ability to properly estimate the covariate distribution decreases. A possible solution consists in increasing the starting sample - assigned via restricted randomization - in order to provide sufficient and reliable information to the algorithm. If instead some a-priori information about the joint distribution of the covariates is available, another possible solution is to rely on standard parametric inference by estimating the parameters of the covariates distribution.

4 | NUMERICAL COMPARISONS

4.1 | The simulated annealing design in the sequential framework

In this section, the operating characteristics of the newly introduced sequential SADe will be compared with several CA rules proposed in the literature. As stated in Remark 2, the dependent structure of the covariates is not relevant for the application of our methodology, so in what follows we will assume independent covariates. More specifically, when all of them are categorical, we will consider:

- stratified randomization performed via the Big Stick Design[37] (BSD), where the maximum tolerated imbalance is set equal to 3, and the Covariate-Adaptive Biased Coin Design[7] (CA-BCD) with allocation function $F(x) = (x^2 + 1)^{-1}, x \geq 1$;

- Pocock and Simon's minimization method (PS) and its generalization by Hu and Hu[6] (HH), with biasing probability set equal to 0.75 and 0.85, respectively;
- Atkinson's D_A-optimum Biased Coin Design (D_A-BCD).

Finally, as a benchmark the completely randomized design (CR), which ignores the information provided by the covariate profile, is also considered. In order to make homogeneous comparisons, we take into account the simulated annealing design in its fully sequential version SADe(1), where we set $T_0 = 50$ and $r = 0.9$, with a total number of 100 temperatures and 200 iterations for each temperature (for a more detailed discussion on how to properly choose parameters values see Section 6). We perform a simulation study where each experiment is simulated 5000 times under four different model specifications:

M1: four binary covariates with only the main effects (without interactions)
M2: four binary covariates with main effects and interactions of all orders (ie, the full model)
M3: ten binary covariates with only the main effects (without interactions)
M4: ten binary covariates with main effects with pairwise interactions

Tables 1 and 2 display the expectations and standard errors of ℓ_n and M_n for all the considered procedures under models M1-M4 as n varies.

In the first scenario (M1), both the minimization procedures and the CA-BCD show good performances, while the BSD and the Atkison's procedure display the highest values. The CR design has the worst performances with imbalance measures that remain stable as the sample size increases: this holds for all the considered scenarios. Even in this simple set-up, with only four covariates, the SADe(1) provides a remarkable gain in terms of the ability to balance the experimental groups. As an example for $n = 50$, the loss of efficiency induced by the SADe(1) is about $0.41/50 = 0.8\%$, that of the CA-BCD about 6.9%, while for the CR design is about 10%. As the complexity of the model specification grows, the improvement of the newly proposed procedure increases: this is particularly evident for M2 and M3. For the M2 and $n = 50$ indeed, the SADe(1) exhibits a loss of efficiency of 9% while for the CA-BCD and CR design about 15% and 31%, respectively. In the last scenario, all the considered procedures struggle to balance the groups, even for $n = 400$, displaying also a high variability of the estimates (note that the case $n = 50$ has been displayed to maintain the same setting throughout the tables; however it is purely demonstrative due to the high number of parameters involved and the quality of multidimensional kernel estimation of $\mathcal{G}$). The SA-based procedure is able to provide quite good performances, especially as the sample size increases; although the D_A-BCD shows slightly better values for intermediate sample sizes, its imbalance measures tend to stabilize as n grows, while for SADe(1) the corresponding measures tend to vanish. In general, SADe(1) guarantees both the lowest loss of information and Mahalanobis distance along with the smallest variability of the estimates, in particular for small/moderate sample sizes. The HH procedure shows good performances in several scenarios but with greater variability with respect to the SA-based procedure and it generally necessitates a higher number of patients to provide well-balanced experimental groups. The PS procedure instead performs quite well as long as the model does not contain interactions (see for example scenario M2), while the BSD and the stratification-based procedures show poor performances when there is a high number of strata wrt the sample size, as in model M3. This is due to the fact that the presence of few statistical units in each stratum prevents these procedures to properly evolve.

For the mixed case of both categorical and continuous covariates, the SADe will be compared to the Atkinson's design and the kernel density-based procedure (KER) proposed by Ma and Hu,[16] where following the authors suggestion the biasing probability is set to 0.80, the same importance is provided to each factor and at each step the covariates are re-scaled. Moreover, we also considered the group sequential rerandomization RR(m) proposed by Zhou et al:[19] as reported in their Supplementary Material, the total expected number of rerandomizations is set to 2000 and each group consists of $m = 50$ experimental units. To properly compare the performances of the new procedure with that of the group sequential rerandomization, we present the results of both the fully sequential SADe(1) and its group sequential version SADe(m) where each group is composed of $m = 50$ experimental units. The results for the completely randomized design are also reported. In this setup, each design is simulated 5000 times and we adopt the following model specifications:

M5: two binary and three normally distributed variables without interactions
M6: two binary and three normally distributed variables with pairwise interactions
M7: four binary and six normally distributed variables without interactions

TABLE 1 Expectations (with standard errors in parentheses) of ℓ_n and M_n for models M1 and M2 as n varies.

		$n = 50$		$n = 100$		$n = 200$		$n = 400$	
	Rules	ℓ_n	M_n	ℓ_n	M_n	ℓ_n	M_n	ℓ_n	M_n
M1	SADe(1)	0.41	0.43	0.25	0.26	0.13	0.14	0.06	0.07
		(0.27)	(0.30)	(0.18)	(0.20)	(0.09)	(0.10)	(0.05)	(0.05)
	PS	2.19	2.03	0.65	0.61	0.29	0.27	0.14	0.13
		(1.83)	(1.72)	(0.58)	(0.55)	(0.24)	(0.24)	(0.12)	(0.12)
	D_A-BCD	1.51	1.20	1.08	0.86	1.02	0.81	1.02	0.80
		(1.00)	(0.89)	(0.67)	(0.60)	(0.64)	(0.57)	(0.64)	(0.57)
	CA-BCD	3.43	2.76	1.80	1.44	0.91	0.73	0.45	0.36
		(2.07)	(1.87)	(1.10)	(1.00)	(0.55)	(0.50)	(0.27)	(0.24)
	BSD	4.19	3.35	2.61	2.09	1.36	1.09	0.67	0.54
		(2.54)	(2.29)	(1.59)	(1.43)	(0.86)	(0.77)	(0.45)	(0.39)
	HH	1.51	1.39	0.43	0.39	0.20	0.18	0.10	0.09
		(1.25)	(1.18)	(0.33)	(0.32)	(0.15)	(0.15)	(0.07)	(0.07)
	CR	4.96	3.95	4.99	4.00	4.97	3.97	4.97	3.97
		(2.99)	(2.67)	(3.00)	(2.71)	(3.16)	(2.84)	(3.11)	(2.76)
M2	SADe(1)	4.64	4.61	3.07	3.06	1.68	1.68	0.88	0.88
		(1.51)	(1.53)	(1.21)	(1.22)	(0.71)	(0.72)	(0.39)	(0.40)
	PS	13.20	12.85	12.13	11.98	11.49	11.41	11.33	11.29
		(4.01)	(3.92)	(4.50)	(4.45)	(4.57)	(4.55)	(4.72)	(4.70)
	D_A-BCD	7.28	6.79	3.81	3.55	3.35	3.13	3.27	3.06
		(2.37)	(2.28)	(1.28)	(1.24)	(1.15)	(1.10)	(1.13)	(1.09)
	CA-BCD	11.64	10.92	6.59	6.21	3.07	2.88	1.46	1.39
		(3.22)	(3.14)	(2.09)	(2.04)	(0.99)	(0.96)	(0.45)	(0.44)
	BSD	13.60	12.73	9.24	8.69	4.64	4.35	2.25	2.11
		(3.64)	(3.55)	(2.67)	(2.59)	(1.54)	(1.48)	(0.98)	(0.93)
	HH	9.77	9.51	4.40	4.32	1.85	1.83	0.88	0.88
		(3.37)	(3.29)	(2.03)	(2.01)	(0.89)	(0.88)	(0.42)	(0.41)
	CR	15.38	14.38	15.92	14.93	16.07	15.07	15.99	14.99
		(4.39)	(4.26)	(5.01)	(4.86)	(5.42)	(5.26)	(5.57)	(5.39)

Table 3 displays the expectations and standard errors of ℓ_n and M_n for models M5-M7 as n varies. The SADe(1) confirms the ability to provide well-balanced treatment groups outperforming all other fully sequential procedures, even better for small samples. The Kernel procedure generally shows high imbalance measures, necessitating at least 200 patients to compete with Atkinson's rule. This is particularly evident for M6 and M7, namely either in the presence of interactions among covariates or when the number of the considered variables increases. The CR design confirms its poor ability to provide covariate balance. As expected, by exploiting more information, the group sequential version SADe(m) provides the best balance outdoing all the considered fully sequential procedures; indeed, SADe(50) shows the lowest values of ℓ_n and M_n with the smallest variability of the estimates in all the considered set-ups. The improvement in the performances of the SADe(m) with respect to the RR(m) procedure is particularly relevant for M6 and M7, namely for more complex scenarios: the imbalance measures provided by the SA-based method are at least one third of those obtained adopting the rerandomization approach.

TABLE 2 Expectations (with standard errors in parentheses) of ℓ_n and M_n for models M3 and M4 as n varies.

		$n = 50$		$n = 100$		$n = 200$		$n = 400$	
	Rules	ℓ_n	M_n	ℓ_n	M_n	ℓ_n	M_n	ℓ_n	M_n
M3	SADe(1)	1.83	1.82	1.22	1.23	0.67	0.67	0.32	0.32
		(0.76)	(0.78)	(0.57)	(0.58)	(0.30)	(0.30)	(0.15)	(0.16)
	PS	7.61	7.38	3.64	3.57	1.69	1.67	0.84	0.83
		(3.49)	(3.40)	(1.92)	(1.89)	(0.92)	(0.92)	(0.47)	(0.46)
	D_A-BCD	4.27	3.85	2.46	2.23	2.27	2.06	2.22	2.01
		(1.91)	(1.79)	(1.03)	(0.98)	(0.96)	(0.91)	(0.93)	(0.89)
	CA-BCD	11.01	10.02	11.01	9.99	10.87	9.88	10.72	9.72
		(4.09)	(3.96)	(4.47)	(4.30)	(4.58)	(4.40)	(4.51)	(4.32)
	BSD	11.03	10.02	10.99	9.98	10.96	9.95	10.95	9.93
		(4.10)	(3.94)	(4.37)	(4.19)	(4.50)	(4.31)	(4.65)	(4.43)
	HH	6.11	5.95	2.08	2.04	0.97	0.95	0.52	0.51
		(3.00)	(2.93)	(1.12)	(1.10)	(0.48)	(0.48)	(0.25)	(0.25)
	CR	11.06	10.06	11.02	10.02	10.95	9.91	10.98	9.95
		(4.07)	(3.93)	(4.42)	(4.21)	(4.56)	(4.32)	(4.69)	(4.44)
M4	SADe(1)	47.80	46.75	39.59	38.96	20.53	20.44	9.64	9.65
		(2.03)	(2.09)	(6.18)	(6.10)	(4.19)	(4.14)	(2.06)	(2.05)
	PS	48.76	47.78	52.54	52.00	49.08	48.82	46.96	46.83
		(1.53)	(1.50)	(7.02)	(6.95)	(8.71)	(8.67)	(8.98)	(8.96)
	D_A-BCD	48.12	47.13	36.14	35.40	17.18	16.83	12.53	12.29
		(1.88)	(1.87)	(5.44)	(5.39)	(3.18)	(3.14)	(2.25)	(2.22)
	CA-BCD	48.83	47.83	56.12	55.11	55.93	54.93	55.08	54.11
		(1.50)	(1.50)	(6.87)	(6.83)	(8.92)	(8.85)	(9.60)	(9.55)
	BSD	48.86	47.86	56.16	55.16	56.12	55.11	55.97	54.99
		(1.50)	(1.50)	(6.99)	(6.96)	(8.91)	(8.84)	(9.83)	(9.75)
	HH	48.46	47.49	50.31	49.79	45.05	44.81	40.12	40.01
		(1.71)	(1.68)	(7.06)	(6.99)	(8.18)	(8.14)	(7.99)	(7.97)
	CR	48.87	47.87	56.01	55.01	55.91	54.88	55.81	54.78
		(1.47)	(1.47)	(6.84)	(6.81)	(8.84)	(8.79)	(9.88)	(9.80)

4.1.1 | The simulated annealing design under the fixed scenario

As an illustrative example of the case in which all the patients' profile are available before the trial starts, we provide the comparison of SADe* with the (fixed) rerandomization procedure, denoted by RR*. Several scenarios are considered: four different sample sizes, $n \in \{50, 100, 200, 400\}$, each with four different number of covariates, namely $q \in \{5, 10, 20, 40\}$. Since the rerandomization has been originally proposed for quantitative variables,[2,19] to properly compare the two procedures, the covariates are assumed to be i.i.d. standard normals. As regards the RR*, the acceptance probability p_a is set equal to 0.001 to enforce a strong covariate balance,[38] while for the SADe* we set $T_0 = 300$ and $r = 0.9$, with a total number of 200 temperatures and 200 iterations for each temperature. Each scenario is simulated 5000 times, both the Mahalanobis distance and the loss of information induced by the two different approaches are considered: Table 4 summarizes the expectation and standard error (in brackets) of ℓ_n and M_n for SADe* and RR* as q and n vary. The imbalance measures provided by RR* grow as q grows and remain stable over the sample sizes: this is due to the fact that the

TABLE 3 Expectations (with standard errors in parentheses) of ℓ_n and M_n for models M5-M7 as n varies.

		$n = 50$		$n = 100$		$n = 200$		$n = 400$	
	Rules	ℓ_n	M_n	ℓ_n	M_n	ℓ_n	M_n	ℓ_n	M_n
M5	SADe(1)	0.57	0.53	0.35	0.33	0.17	0.16	0.08	0.07
		(0.34)	(0.32)	(0.23)	(0.22)	(0.11)	(0.11)	(0.05)	(0.05)
	SADe(50)	0.08	0.06	0.04	0.03	0.02	0.01	0.01	0.01
		(0.04)	(0.04)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	D_A-BCD	1.96	1.63	1.29	1.06	1.23	1.03	1.22	1.01
		(1.18)	(1.07)	(0.74)	(0.67)	(0.71)	(0.65)	(0.69)	(0.63)
	KER	3.35	3.20	1.82	1.77	1.18	1.16	0.78	0.76
		(2.23)	(2.17)	(1.33)	(1.31)	(0.88)	(0.87)	(0.59)	(0.59)
	RR(50)	0.13	0.13	0.08	0.08	0.04	0.04	0.02	0.02
		(0.04)	(0.04)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
	CR	6.10	5.12	6.04	5.04	6.07	5.06	6.07	5.03
		(3.26)	(3.05)	(3.36)	(3.10)	(3.43)	(3.14)	(3.44)	(3.15)
M6	SADe(1)	3.20	3.40	2.47	2.63	1.71	1.81	1.23	1.29
		(1.19)	(1.30)	(1.17)	(1.25)	(0.95)	(0.99)	(0.73)	(0.76)
	SADe(50)	0.89	0.83	0.41	0.38	0.19	0.18	0.09	0.09
		(0.26)	(0.25)	(0.12)	(0.12)	(0.06)	(0.06)	(0.03)	(0.03)
	D_A-BCD	8.34	7.80	4.27	3.99	3.53	3.31	3.36	3.14
		(2.47)	(2.64)	(1.47)	(1.42)	(1.19)	(1.16)	(1.17)	(1.13)
	KER	13.94	13.60	12.22	12.07	11.45	11.37	10.95	10.91
		(4.14)	(4.06)	(4.31)	(4.27)	(4.46)	(4.43)	(4.55)	(4.54)
	RR(50)	2.65	2.60	1.82	1.80	0.81	0.80	0.40	0.40
		(0.45)	(0.44)	(0.45)	(0.44)	(0.19)	(0.19)	(0.09)	(0.09)
	CR	16.03	15.05	15.99	14.98	16.01	15.00	16.00	14.96
		(4.39)	(4.28)	(5.00)	(4.84)	(5.30)	(5.14)	(5.44)	(5.26)
M7	SADe(1)	1.81	1.77	1.25	1.23	0.65	0.64	0.32	0.31
		(0.75)	(0.75)	(0.58)	(0.56)	(0.32)	(0.32)	(0.15)	(0.14)
	SADe(50)	0.39	0.35	0.19	0.17	0.09	0.08	0.05	0.04
		(0.10)	(0.10)	(0.05)	(0.05)	(0.02)	(0.03)	(0.01)	(0.01)
	D_A-BCD	4.34	3.90	2.47	2.23	2.30	2.08	2.22	2.02
		(1.91)	(1.80)	(1.02)	(0.98)	(0.97)	(0.93)	(0.96)	(0.91)
	KER	7.92	7.70	4.97	4.89	3.40	3.36	2.38	2.36
		(3.43)	(3.35)	(2.42)	(2.39)	(1.73)	(1.72)	(1.23)	(1.23)
	RR(50)	1.06	1.04	0.82	0.81	0.34	0.34	0.18	0.18
		(0.19)	(0.19)	(0.17)	(0.16)	(0.06)	(0.06)	(0.03)	(0.03)
	CR	11.06	10.05	11.05	10.02	11.06	10.02	11.06	10.02
		(4.12)	(3.97)	(4.48)	(4.28)	(4.64)	(4.40)	(4.61)	(4.38)

TABLE 4 Expectations (with standard errors in parentheses) of ℓ_n and M_n for SADe* and RR* as q and n vary.

		$n = 50$		$n = 100$		$n = 200$		$n = 400$	
q	Rules	ℓ_n	M_n	ℓ_n	M_n	ℓ_n	M_n	ℓ_n	M_n
5	SADe*	0.04	0.04	0.03	0.03	0.02	0.02	0.01	0.01
		(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
	RR*	1.14	0.15	1.14	0.15	1.18	0.15	1.16	0.15
		(1.40)	(0.05)	(1.38)	(0.05)	(1.48)	(0.05)	(1.43)	(0.05)
10	SADe*	0.33	0.29	0.16	0.14	0.08	0.07	0.04	0.04
		(0.09)	(0.09)	(0.04)	(0.04)	(0.02)	(0.02)	(0.01)	(0.01)
	RR*	2.22	1.22	2.17	1.21	2.22	1.21	2.21	1.21
		(1.37)	(0.22)	(1.38)	(0.22)	(1.39)	(0.22)	(1.44)	(0.22)
20	SADe*	1.59	1.49	0.72	0.68	0.38	0.36	0.20	0.19
		(0.31)	(0.31)	(0.15)	(0.15)	(0.08)	(0.08)	(0.04)	(0.04)
	RR*	6.29	5.29	6.25	5.26	6.22	5.25	6.24	5.24
		(1.37)	(0.55)	(1.41)	(0.57)	(1.47)	(0.59)	(1.54)	(0.59)
40	SADe*	8.51	8.18	3.30	3.19	1.80	1.74	1.01	0.98
		(1.16)	(1.14)	(0.49)	(0.49)	(0.29)	(0.29)	(0.18)	(0.18)
	RR*	17.99	16.99	17.76	16.77	17.67	16.69	17.61	16.65
		(1.28)	(0.89)	(1.57)	(1.06)	(1.69)	(1.10)	(1.72)	(1.14)

acceptance probability is directly related to the number of covariates and not to the sample size. The SADe* instead guarantees monotonically decreasing and less variable imbalance measures as the number of considered subjects increases; on average, the values for $n = 400$ are about one eighth of that for $n = 50$. In general, the SADe* provides a consistent improvement wrt RR in every scenario, as an example for $q = 5$ and $n = 400$ the loss induced by the adoption of the newly proposed method is more than 100 times smaller than that of RR*, while the Mahalanobis distance is about one fifteenth that obtained via rerandomization. While for the SADe* the values of ℓ_n and M_n tend to be close, for RR* this seems to hold only when $q > 10$.

4.2 | Case study

In this section, we apply our proposed methodology to redesign a real trial, by using the clinical data of The Cancer Genome Atlas Uterine Corpus Endometrial Carcinoma (TCGA-UCEC) project,[39,40] freely available at NCI Genomic Data Commons—https://gdc.cancer.gov/. From the original dataset, consisting of clinical and demographic information of $n = 548$ UCEC cases, we select eleven covariates, two continuous and nine categorical, that have few missing values and are most likely associated with the severity of tumor symptoms. Categorical/ordinal covariates are subdivided into dummy variables according to their different levels: this translates into a total of 40 variables considered. The missing values are imputed by sampling from the observed values. Six different procedures are compared: the SADe (both group and fully sequential), the group sequential rerandomization, the Atkinson's optimum design, the Kernel procedures and the completely randomized design; each design is simulated 5000 times. For both RR(m) and SADe(m), the enrollment structure consists in $k = 10$ groups: the first 6 of them are composed of 54 patients and the remaining 4 of 56 units (as reported in design (v) pag. 11 of the Supplementary Material of Zhou et al[19]).

The results in Table 5 confirm the findings of the previous section where the CR design showed the worst performances. The Kernel procedure, the second worst, exhibits imbalance measures that are about 2.5 times those the D_A-BCD; this is probably due to the high number of covariates considered (see for example M6 in Table 3 of the previous section). The fully sequential SADe exhibits values of the imbalance measures that are smaller than those of the D_A-BCD but slightly more variable; as long as the group sequential designs are considered, our proposal outperforms all the other procedures both in terms of balance and variability of the estimates. In particular, if compared with the RR(m), the

TABLE 5 Expectations (with standard errors in parentheses) of ℓ_n and M_n for TCGA-UCEC data.

Rules	ℓ_n		M_n	
SADe(1)	8.27	(2.22)	8.27	(2.21)
SADe(m)	4.57	(0.63)	4.57	(0.63)
D_A-BCD	8.98	(1.68)	8.75	(1.65)
KER	20.18	(4.42)	20.14	(4.41)
RR(m)	7.27	(1.44)	7.25	(1.44)
CR	28.96	(6.30)	27.94	(6.15)

SA-based procedure exhibits a reduction of more than 35% in both the induced loss of information and the Mahalanobis distance.

5 | THE IMPACT OF BALANCING COVARIATES ON THE INFERENTIAL PRECISION

The aim of this section is to stress the practical importance of balancing covariates over the experimental groups in guaranteeing a solid statistical inference about the treatment effects. In what follows, all the considered procedures will be compared in terms of the estimation accuracy of the treatment difference $\theta_A - \theta_B$ and the statistical power. More specifically, based on model (1), the patients' responses are assumed to be normally distributed with different treatment effects, θ_A and θ_B, and common variance $\sigma^2 = 1$; each design is simulated 5000 times and the covariate effects are set to be $\beta_j = 1, j = 1, \ldots p$. Under the same simulation setting of Section 4.1, Table 6 shows the average power and the estimated treatment difference for models M1-M3, for $\theta_A - \theta_B = 0.3$ as n varies.

Several conclusions can be drawn from these results. In general, the behavior of the considered designs in terms of both statistical power and estimation accuracy resembles that found for imbalance measures in the previous sections. While in almost all the cases and for most of the considered procedures the estimated treatment difference is close to the true value of 0.3, its variability greatly varies across the designs. The CR design always displays the highest standard errors and the lowest power, while the SADe design exhibits the highest power along with the lowest variability of the estimates, especially for small and moderate sample sizes and for more complex model specifications. Indeed, considering M3 with $n = 50$ for example, the gain in power with respect to the competitors is at least of 3% with a reduction in the variability of the estimates of about 7%. Similar conclusions can be drawn for $n = 100$, clearly as n increases the impact of the covariate imbalance on the inferential precision tends to be mitigated. The HH procedure confirms good performances, with values close to the proposed procedure but with a higher variability in the estimates. The results about CA-BCD and the BSD confirm their weaknesses in the presence of a high number of variables, while PS procedure shows low power when the model contains interaction terms. For what concerns the mixed case of both qualitative and quantitative variables, Table 7 summarizes the average power and estimated treatment difference for models M5-M7, for $\theta_A - \theta_B = 0.3$ as n varies. The results strongly emphasize the importance of adopting an adaptive procedure that forces covariate balance: this is particularly evident looking at the poor performances of the CR design. The good behavior of SADe(1) in terms of inferential precision is confirmed: it displays the highest power and the lowest variability among all the fully sequential procedures, with values close to those of RR(50). In general, the D_A-BCD performs better than KER especially for more complex models and for grater sample sizes. Finally, the group sequential version, SADe(50), guarantees the overall best estimation precision and statistical power.

6 | DISCUSSION

In this article, we propose a new procedure based on the Simulated Annealing algorithm aimed at balancing the covariates of the experimental groups. The SADe turns out to be very flexible with several points of strength: (i) it can deal with any kind of covariate (either qualitative or quantitative), any model specification and any imbalance measure; (ii) it can be easily adapted for fixed, group or fully sequential experiments; (iii) it can also be easily accommodated for multi-arm

TABLE 6 Power and estimation accuracy comparison for models M1-M3 and $\theta_A - \theta_B = 0.3$ as n varies (with standard errors in parentheses).

		M1		M2		M3	
n	Rules	Power	$\hat{\theta}_{An} - \hat{\theta}_{Bn}$	Power	$\hat{\theta}_{An} - \hat{\theta}_{Bn}$	Power	$\hat{\theta}_{An} - \hat{\theta}_{Bn}$
50	SADe(1)	0.21	0.30 (0.28)	0.19	0.30 (0.30)	0.20	0.30 (0.28)
	PS	0.19	0.31 (0.30)	0.16	0.29 (0.33)	0.16	0.29 (0.31)
	D_A-BCD	0.19	0.30 (0.30)	0.17	0.30 (0.31)	0.17	0.30 (0.32)
	CA-BCD	0.18	0.31 (0.30)	0.16	0.32 (0.33)	0.15	0.30 (0.33)
	BSD	0.18	0.31 (0.30)	0.16	0.30 (0.33)	0.15	0.30 (0.33)
	HH	0.19	0.30 (0.30)	0.17	0.31 (0.32)	0.16	0.30 (0.30)
	CR	0.14	0.30 (0.31)	0.13	0.29 (0.34)	0.14	0.31 (0.33)
100	SADe(1)	0.34	0.30 (0.19)	0.32	0.30 (0.20)	0.33	0.30 (0.19)
	PS	0.32	0.30 (0.21)	0.28	0.30 (0.22)	0.31	0.30 (0.21)
	D_A-BCD	0.32	0.30 (0.21)	0.31	0.29 (0.21)	0.32	0.30 (0.20)
	CA-BCD	0.31	0.31 (0.21)	0.30	0.31 (0.21)	0.29	0.30 (0.21)
	BSD	0.31	0.29 (0.21)	0.30	0.30 (0.21)	0.29	0.29 (0.21)
	HH	0.33	0.30 (0.20)	0.31	0.30 (0.21)	0.32	0.30 (0.20)
	CR	0.24	0.30 (0.22)	0.23	0.29 (0.22)	0.24	0.31 (0.22)
200	SADe(1)	0.58	0.30 (0.13)	0.57	0.30 (0.14)	0.56	0.30 (0.13)
	PS	0.57	0.30 (0.14)	0.53	0.30 (0.15)	0.55	0.30 (0.14)
	D_A-BCD	0.56	0.31 (0.15)	0.55	0.30 (0.14)	0.55	0.30 (0.15)
	CA-BCD	0.56	0.30 (0.14)	0.56	0.30 (0.14)	0.53	0.30 (0.15)
	BSD	0.56	0.30 (0.15)	0.55	0.30 (0.14)	0.53	0.30 (0.15)
	HH	0.58	0.30 (0.14)	0.57	0.30 (0.14)	0.56	0.30 (0.14)
	CR	0.44	0.31 (0.15)	0.43	0.30 (0.15)	0.44	0.30 (0.15)
400	SADe(1)	0.87	0.30 (0.10)	0.86	0.30 (0.10)	0.86	0.30 (0.10)
	PS	0.86	0.30 (0.10)	0.84	0.30 (0.10)	0.85	0.30 (0.10)
	D_A-BCD	0.85	0.30 (0.10)	0.85	0.30 (0.10)	0.84	0.30 (0.10)
	CA-BCD	0.86	0.31 (0.10)	0.85	0.30 (0.10)	0.83	0.30 (0.10)
	BSD	0.86	0.30 (0.10)	0.84	0.30 (0.10)	0.83	0.30 (0.10)
	HH	0.87	0.30 (0.10)	0.86	0.30 (0.10)	0.86	0.30 (0.10)
	CR	0.76	0.30 (0.11)	0.76	0.30 (0.10)	0.76	0.30 (0.10)

experiments by choosing a suitable imbalance measure. Moreover, thanks to the stochastic nature of the SA algorithm, which under widely satisfied conditions converges to the global optimum, the SADe procedure is completely unpredictable. An extensive simulation study has shown the good performances of the new design which can outperform all the procedures already presented in the literature, in particular for small/moderate sample sizes and in the presence of several covariates.

When implementing SA-based procedures, two important aspects are the cooling scheme and the neighbors specification. For what concerns the first one, although there might be an optimal scheme for each specific problem, a general valid solution is the standard geometric scheme: this has the advantage of being robust and guarantees convergence,[29,32,33] but it may require more time to converge (than it would do with an optimal cooling scheme). In any case, our results (not shown here for sake of brevity) indicate that the performances as well as the computational time are not greatly affected by the chosen cooling scheme. Moreover, the SADe designs turns out to be particularly robust in terms of the choice of both the decay parameter r and the initial temperature T_0. Indeed, our evidence shows that the value of r does not greatly

TABLE 7 Power and estimation accuracy comparison for models M5-M7 and $\theta_A - \theta_B = 0.3$ as n varies (with standard errors in parentheses).

		M5		M6		M7	
n	Rules	Power	$\hat{\theta}_{An} - \hat{\theta}_{Bn}$	Power	$\hat{\theta}_{An} - \hat{\theta}_{Bn}$	Power	$\hat{\theta}_{An} - \hat{\theta}_{Bn}$
50	SADe(1)	0.21	0.30 (0.28)	0.21	0.30 (0.29)	0.21	0.30 (0.29)
	SADe(50)	0.22	0.30 (0.28)	0.22	0.30 (0.28)	0.22	0.30 (0.28)
	D_A-BCD	0.19	0.30 (0.29)	0.17	0.30 (0.31)	0.18	0.31 (0.31)
	KER	0.17	0.29 (0.30)	0.16	0.31 (0.34)	0.16	0.28 (0.31)
	RR(50)	0.22	0.31 (0.28)	0.21	0.30 (0.30)	0.21	0.30 (0.29)
	CR	0.14	0.30 (0.31)	0.13	0.30 (0.34)	0.13	0.29 (0.32)
100	SADe(1)	0.33	0.30 (0.19)	0.32	0.30 (0.20)	0.33	0.30 (0.20)
	SADe(50)	0.34	0.30 (0.19)	0.34	0.30 (0.19)	0.34	0.30 (0.19)
	D_A-BCD	0.31	0.30 (0.20)	0.31	0.31 (0.21)	0.32	0.30 (0.20)
	KER	0.31	0.30 (0.20)	0.28	0.30 (0.22)	0.29	0.29 (0.20)
	RR(50)	0.34	0.31 (0.19)	0.33	0.30 (0.19)	0.33	0.30 (0.20)
	CR	0.23	0.30 (0.21)	0.22	0.30 (0.22)	0.23	0.30 (0.22)
200	SADe(1)	0.58	0.30 (0.13)	0.57	0.30 (0.13)	0.57	0.30 (0.13)
	SADe(50)	0.58	0.30 (0.12)	0.58	0.30 (0.13)	0.58	0.30 (0.12)
	D_A-BCD	0.56	0.30 (0.14)	0.55	0.30 (0.14)	0.55	0.30 (0.14)
	KER	0.56	0.30 (0.15)	0.53	0.30 (0.15)	0.53	0.30 (0.14)
	RR(50)	0.58	0.30 (0.14)	0.57	0.30 (0.14)	0.57	0.30 (0.13)
	CR	0.45	0.30 (0.15)	0.43	0.30 (0.15)	0.44	0.30 (0.15)
400	SADe(1)	0.87	0.30 (0.10)	0.86	0.30 (0.10)	0.86	0.30 (0.10)
	SADe(50)	0.87	0.30 (0.09)	0.87	0.30 (0.09)	0.87	0.30 (0.09)
	D_A-BCD	0.85	0.30 (0.10)	0.85	0.30 (0.10)	0.85	0.30 (0.10)
	KER	0.85	0.30 (0.11)	0.83	0.30 (0.11)	0.85	0.30 (0.11)
	RR(50)	0.87	0.30 (0.10)	0.86	0.30 (0.10)	0.86	0.30 (0.10)
	CR	0.75	0.30 (0.11)	0.75	0.30 (0.11)	0.75	0.30 (0.11)

affect the performance of SADe: we suggest a value of $r = 0.9$ and 200 iterations per temperature as they represent a good compromise between computational time and ability to well balance the experimental group for all the considered scenarios. As an example, by taking into account the model specification M2 with $n = 100$ and setting $r = 0.80$ we obtain $\ell_n = 3.07$ and $M_n = 3.08$, while by choosing $r = 0.99$ the results are $\ell_n = 3.10$ and $M_n = 3.12$, which are really close to the values obtained with $r = 0.9$, namely $\ell_n = 3.07$ and $M_n = 3.06$ (see Table 1). As another example, let us consider M7 with $n = 100$, by setting $r = 0.8$ we get $\ell_n = 1.24$ and $M_n = 1.23$, whereas with $r = 0.99$ the figures are $\ell_n = 1.29$ and $M_n = 1.28$ that should be compared with those reported in Table 3, namely $\ell_n = 1.25$ and $M_n = 1.23$. For what concerns the initial temperature T_0, as shown in Section 4, a value of 300 for the fixed case and of 50 for the sequential one can be valid starting points, since they provide good results for a wide variety of scenarios; clearly, as for the decay parameter, these values can be further tuned for any case-specific setting through a grid search. Nevertheless, it is worth stressing that the proposed procedure is also robust in terms of the chosen initial temperature: as an example, consider again the model specification M2 with $n = 100$, a value of $T_0 = 10$ leads to $\ell_n = 3.08$ and $M_n = 3.06$, while by setting $T_0 = 250$ we obtain $\ell_n = 3.06$ and $M_n = 3.08$, values that are practically indistinguishable from those achieved with $T_0 = 50$ (see Table 1). By considering M7 with $n = 100$ instead, by choosing $T_0 = 10$ the results are $\ell_n = 1.23$ and $M_n = 1.22$, while with $T_0 = 250$ the imbalance measures are $\ell_n = 1.27$ and $M_n = 1.25$, again values very close to those obtained with $T_0 = 50$ (see Table 3).

The second aspect concerns the neighbors specification. Here we adopt a variable neighborhood search[34-36] in which the search space gradually narrows as the iterations progress. For this particular kind of problem, such strategy seems

to be preferable, outdoing the fixed neighbors search case. In order to facilitate the implementation of our proposed strategy, also with different values of the tuning parameters and/or a different imbalance measure, an example code and instructions for running it are available in the online Supplementary Material.

As a possible extension of the present work, an interesting new line of research could focus on the definition of a suitable imbalance metric that enables to incorporate the 'ethical' goal of assigning more patients to the best performing treatment. This would allow to allocate patients by considering both their covariate profiles and the responses to the treatment, as is the case for covariate-adjusted response-adaptive (CARA) designs:[1] in this setting indeed, the optimal allocation may greatly differ from the balanced one depending on the effectiveness of the considered treatments. Another possible extension regards the generalization of the proposed procedure to $L \geq 2$ multiple group design problems. This can be done by properly modifying the imbalance measures of interest: for what concerns the loss of information an example can be found in Atkinson,[12] while the Mahalanobis distance can be modified as $M_n = \sum_{l=1}^{L} \pi_l (\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}})^t \mathrm{var}(\boldsymbol{x})^{-1} (\bar{\boldsymbol{x}}_l - \bar{\boldsymbol{x}})$ where π_l represents the percentage of allocation to the l-group, $\bar{\boldsymbol{x}}_l$ and $\bar{\boldsymbol{x}}$ are the sample means for the l-group and for all the observations, respectively.

DATA AVAILABILITY STATEMENT

The data that supports the finding of this study are freely available at NCI Genomic Data Commons—https://gdc.cancer.gov/.

ACKNOWLEDGMENT

Open Access Funding provided by Universitá degli Studi di Bologna within the CRUI-CARE Agreement.

ORCID

Alessandro Baldi Antognini https://orcid.org/0000-0002-7426-8208
Marco Novelli https://orcid.org/0000-0002-8827-7797
Maroussa Zagoraiou https://orcid.org/0000-0002-8461-8240

REFERENCES

1. Rosenberger WF, Sverdlov O. Handling covariates in the design of clinical trials. *Stat Sci*. 2008;23:404-419.
2. Morgan KL, Rubin DB. Rerandomization to improve covariate balance in experiments. *Ann Stat*. 2012;40(2):1263-1282.
3. Hu F, Hu Y, Ma Z, Rosenberger WF. Adaptive randomization for balancing over covariates. *Wiley Interdiscip Rev Comput Stat*. 2014;6(4):288-303.
4. Taves DR. Minimization: a new method of assigning patients to treatment and control groups. *J Clin Pharm Ther*. 1974;15:443-453.
5. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*. 1975;31:103-115.
6. Hu Y, Hu F. Asymptotic properties of covariate-adaptive randomization. *Ann Stat*. 2012;40:1794-1815.
7. Baldi Antognini A, Zagoraiou M. The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika*. 2011;98:519-535.
8. Ciolino J, Zhao W, Palesch Y, et al. Quantifying the cost in power of ignoring continuous covariate imbalances in clinical trial randomization. *Contemp Clin Trials*. 2011;32(2):250-259.
9. Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials: a review. *Control Clin Trials*. 2002;23(6):662-674.
10. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006;25(1):127-141.
11. Williamson SF, Villar SS. A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. *Biometrics*. 2020;76(1):197-209.
12. Atkinson AC. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*. 1982;69:61-67.
13. Smith RL. Properties of biased coin designs in sequential clinical trials. *Ann Stat*. 1984;12:1018-1034.
14. Smith RL. Sequential treatment allocation using biased coin designs. *J R Stat Soc Ser B*. 1984;46:519-543.
15. Zagoraiou M. Choosing a covariate-adaptive randomization procedure in practice. *J Biopharm Stat*. 2017;27(5):845-857.
16. Ma Z, Hu F. Balancing continuous covariates based on kernel densities. *Contemp Clin Trials*. 2013;34(2):262-269.
17. Rosenberger WF, Lachin JM. *Randomization in clinical trials: theory and practice*. New York: John Wiley & Sons; 2015.
18. Baldi Antognini A, Giovagnoli A. *Adaptive Designs for Sequential Treatment Allocation*. New York: Chapman & Hall/CRC Biostatistics; 2015.
19. Zhou Q, Ernst PA, Morgan KL, Rubin DB, Zhang A. Sequential rerandomization. *Biometrika*. 2018;105(3):745-752.
20. Khan O, Fotheringham S, Wood V, et al. HR23B is a biomarker for tumor sensitivity to HDAC inhibitor-based therapy. *Proc Natl Acad Sci*. 2010;107(14):6532-6537.

21. Li Y, Sheu CC, Ye Y, et al. Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *Lancet Oncol*. 2010;11(4):321-330.
22. Ashley EA, Butte AJ, Wheeler MT, et al. Clinical assessment incorporating a personal genome. *Lancet*. 2010;375(9725):1525-1535.
23. Hu F. Statistical issues in trial design and personalized medicine. *Clin Investig (Lond)*. 2012;2(2):121-124.
24. Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet*. 2018;19(5):299.
25. Rudrapatna VA, Butte AJ, et al. Opportunities and challenges in using real-world data for health care. *J Clin Invest*. 2020;130(2):565-574.
26. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys*. 1953;21(6):1087-1092.
27. Wolsey LA, Nemhauser GL. *Integer and combinatorial optimization*. Vol 55. New York: John Wiley & Sons; 1999.
28. Aarts E, Korst J, Michiels W. Simulated annealing. In *Search methodologies*. Boston, MA: Springer; 2005;187-210.
29. Gendreau M, Potvin JY, et al. *Handbook of metaheuristics. 2*. Cham: Springer; 2010.
30. Van Laarhoven PJM, Aarts EHL. *Simulated annealing: Theory and applications*. Netherlands: Springer; 1987.
31. Bertsimas D, Tsitsiklis J. Simulated annealing. *Stat Sci*. 1993;8(1):10-15.
32. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. *Science*. 1983;220(4598):671-680.
33. Johnson DS, Aragon CR, McGeoch LA, Schevon C. Optimization by simulated annealing: An experimental evaluation; part I, graph partitioning. *Oper Res*. 1989;37(6):865-892.
34. Mladenović N, Hansen P. Variable neighborhood search. *Comput Oper Res*. 1997;24(11):1097-1100.
35. Bouffard V, Ferland JA. Improving simulated annealing with variable neighborhood search to solve the resource-constrained scheduling problem. *J Sched*. 2007;10(6):375-386.
36. Palubeckis G. A variable neighborhood search and simulated annealing hybrid for the profile minimization problem. *Comput Oper Res*. 2017;87:83-97.
37. Soares JF, Wu CFJ. Some restricted randomization rules in sequential designs. *Commun Stat Theory Methods*. 1983;12:2017-2034.
38. Morgan KL, Rubin DB. Rerandomization to balance tiers of covariates. *J Am Stat Assoc*. 2015;110(512):1412-1421.
39. Levine DA. Integrated genomic characterization of endometrial carcinoma. *Nature*. 2013;497(7447):67-73.
40. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. *N Engl J Med*. 2016;375(12):1109-1112.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Baldi Antognini A, Novelli M, Zagoraiou M. Simulated annealing for balancing covariates. *Statistics in Medicine*. 2023;1-15. doi: 10.1002/sim.9672