TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

Minimization in randomized clinical trials

Elisabeth Coart[1] | **Perrine Bamps[1]** | **Emmanuel Quinaux[1]** |
Geneviève Sturbois[1] | **Everardo D. Saad[1]** | **Tomasz Burzykowski[1,2]** | **Marc Buyse[1,2]**

[1]IDDI, Louvain-la-Neuve, Belgium

[2]Data Science Institute, Hasselt University, Hasselt, Belgium

Correspondence
Elisabeth Coart, IDDI (International Drug Development Institute), 30 avenue Provinciale, 1340 Louvain-la-Neuve, Belgium.
Email: elisabeth.coart@iddi.com

In randomized trials, comparability of the treatment groups is ensured through allocation of treatments using a mechanism that involves some random element, thus controlling for confounding of the treatment effect. Completely random allocation ensures comparability between the treatment groups for all known and unknown prognostic factors. For a specific trial, however, imbalances in prognostic factors among the treatment groups may occur. Although accidental bias can be avoided in the presence of such imbalances by stratifying the analysis, most trialists, regulatory agencies, and other stakeholders prefer a balanced distribution of prognostic factors across the treatment groups. Some procedures attempt to achieve balance in baseline covariates, by stratifying the allocation for these covariates, or by dynamically adapting the allocation using covariate information during the trial (covariate-adaptive procedures). In this Tutorial, the performance of minimization, a popular covariate-adaptive procedure, is compared with two other commonly used procedures, completely random allocation and stratified blocked designs. Using individual patient data of 2 clinical trials (in advanced ovarian cancer and age-related macular degeneration), the procedures are compared in terms of operating characteristics (using asymptotic and randomization tests), predictability of treatment allocation, and achieved balance. Fifty actual trials of various sizes that applied minimization for treatment allocation are used to investigate the achieved balance. Implementation issues of minimization are described. Minimization procedures are useful in all trials but especially when (1) many major prognostic factors are known, (2) many centers of different sizes accrue patients, or (3) the trial sample size is moderate.

1 | INTRODUCTION

Estimation and testing of differences between treatments are the main objectives of randomized clinical trials. In randomized trials, comparability of the treatment groups is ensured through allocation of the treatments under investigation to patients using a mechanism that involves some random element, thus controlling for confounding of the treatment effect.[1] Many randomization procedures have been proposed in the literature and implemented in practice. Completely random allocation, also known as simple randomization, assigns a treatment to every new patient without regard to baseline factors known to be of prognostic importance. Completely random allocation ensures comparability between the treatment groups with respect to all known and unknown prognostic factors in a systematic way.[2] For a

specific trial, however, completely random allocation may result in imbalances, whereby a prognostic factor is unequally distributed between the various treatment groups. Likewise, there may be deviations from the target allocation ratio of patients randomized to the different groups. Some procedures attempt to prevent such imbalances and deviations, either by modifying the probability of treatment allocation over the course of the trial (biased-coin and urn designs), or by restricting randomization (truncated binomial and permuted-block designs). Other procedures attempt to achieve balance between the treatment groups with respect to baseline covariates, either by stratifying the allocation for these covariates, or by dynamically adapting the allocation using covariate information over the course of the trial. The latter set of procedures, known as covariate-adaptive, is the focus of the present Tutorial. More specifically, we compare a popular covariate-adaptive procedure, known as minimization, with two other commonly used procedures, completely random allocation and stratified blocked designs (permuted blocks within strata).

The purpose of minimization is to safeguard against accidental bias with respect to several prognostic factors considered simultaneously. As such, it allows trialists to choose among all baseline factors known or suspected to have a prognostic impact on the outcome of interest, which can then be balanced in the treatment-allocation procedure.[3] Minimization has been shown to result in small imbalances between the groups with respect to the variables used in the allocation process even in trials of small size and/or when the number of prognostic factors is large.[4] Minimization is a valid treatment allocation procedure,[5] yet regulatory guidelines have generally called for caution when minimization is used, especially when a deterministic algorithm is employed.[6] Some have questioned the interest of balancing treatment groups for prognostic factors and have pointed out that such balance does not meet any optimality criterion.[7,8] Others have suggested that minimization may increase the degree of foreknowledge of treatment allocations.[9] Technical difficulties have been documented when minimization is used for unequal allocation ratios.[10] An important concern relates to the appropriate analysis following the use of dynamic allocation. Most trialists use conventional asymptotic tests that have been shown to control the type-I error probability when minimization was used but can be conservative when the test is not stratified.[11-14] It is often suggested that keeping with the famous admonition, "*analyze as you randomize*", attributed to R. A. Fisher, randomization tests should preferably be used to reflect the allocation procedure.[15,16] The type-I error probability and power of such randomization tests, as compared with their asymptotic counterparts, then come into question.

Despite alleged concerns about minimization, it is a widely used treatment allocation procedure, whether in trials conducted by cooperative groups or in pivotal trials for registration of new drugs conducted by the pharmaceutical industry.[17] Some of the most successful drugs ever approved showed their efficacy in trials that allocated treatments to patients using minimization.[18] Minimization has been used in trials of all sizes, ranging from a few dozen patients[19] to tens of thousands of patients.[20] The objective of the present paper is to describe the minimization procedure, address the potential concerns about its use, and study its operating characteristics using simulations and data from actual clinical trials. This Tutorial extends previously published reviews of the properties of treatment allocation procedures, including completely random allocation,[21] stratified blocked designs,[22] adaptive randomization,[23] urn designs,[24] and minimization.[25,26] Of note, a recent review of different treatment-allocation procedures did not cover minimization.[27]

Section 2 sets the scene and introduces terminology for the three most commonly used treatment-allocation procedures in comparative clinical trials, that is, completely random allocation, stratified blocked allocation, and minimization. For the sake of simplicity, we cover simple versions of these procedures, leaving aside extensions that may outperform the simpler versions at the cost of increased complexity.[28-30] Readers who are unfamiliar with randomization techniques will find an excellent coverage of this topic in the classic books by Rosenberger and Lachin,[1] and by Berger.[31] Section 3 presents the design of a clinical trial for patients with advanced ovarian cancer to provide insight into the operating characteristics of the various treatment allocation procedures in terms of balance achieved, type-I error probability and power of the statistical tests used to compare treatment groups, and predictability of treatment allocations. This section further includes an evaluation of the balance achieved in 50 actual trials of various sizes that used minimization. Section 4 presents the analysis of a clinical trial for patients with age-related macular degeneration (AMD) to compare the properties of asymptotic tests vs randomization tests following dynamic treatment allocation. Section 5 considers implementation issues when minimization is the method of choice and covers reporting issues in general. Section 6 concludes the paper with a discussion of the topics to consider when choosing a treatment-allocation procedure for a randomized clinical trial.

2 | BACKGROUND

2.1 | Terminology

2.1.1 | Baseline covariates and factors

Baseline covariates may include patient characteristics (age, sex, ethnicity, performance status, co-morbidities, etc), disease characteristics (severity, stage, location, time since diagnosis, etc), or environmental characteristics (equipment available and other external factors having a potential impact on the patient, disease, or treatment). Baseline covariates that have prognostic importance are usually considered for different forms of stratification and commonly referred to as "factors". When baseline factors are measured on a continuous scale, they are typically categorized for the purposes of randomization (eg, age <50 vs ≥50 years), although minimization can be extended to directly handle continuous variables. In this way, information loss due to discretization can be avoided as well as uncertainties around determining cut-off points.[26,32] The different categories of a baseline factor are referred to as factor levels.

Baseline factors are sometimes known unreliably at the time of randomization: for instance, disease stage is often verified by an independent committee who may later modify the investigator's assessment. In such instances, we shall assume that the information provided at the time of randomization will not be corrected, so that any subsequent modifications to baseline data are captured distinctly from the data used for the treatment allocation (whether using minimization or any other procedure based on baseline data). This approach is required to avoid manipulating baseline data with a view to biasing future allocations; for instance, if poor-prognosis patients in the control group were modified to appear as good-prognosis patients, any balancing scheme would tend to allocate future poor-prognosis patients preferentially to the control group. For the sake of assessing the balance attained, one would use the information provided at the time of randomization, as the resulting allocations may otherwise look imbalanced in terms of the amended factor classifications. Keeping both the data known at the time of randomization and the correct data known at some later time point is also needed to conduct different analyses; specifically, the former should be used to carry out stratified analyses, while the latter would be a more sensible choice for subgroup analyses.

2.1.2 | Centers

The center (or investigational site) is a hospital, clinic, physician, or a group of physicians that can be considered as a unit. Center is a baseline factor that deserves special consideration, for logistic reasons (eg, when drug supply depends on the treatment allocation) and analytical reasons (eg, when center effects are thought to exist, or need to be tested). Sometimes centers accrue very few patients each, in which case they are sometimes grouped to form bigger units at the level of a region, a country, or a continent. The center is known at baseline but can subsequently change, for example, if patients move or decide for some other reason to be followed up in a different center than the one in which they were randomized. In this paper, such changes over the course of the trial will be ignored.

2.1.3 | Stratification

Stratification consists of defining nonoverlapping patient subsets based on some measurable characteristics at baseline, usually with cross-classification of such subsets leading to strata. For instance, stratification for sex and age (<50 vs ≥50 years) would define four distinct strata. In general, a stratification based on N factors having l_i levels ($i = 1, \ldots, N$) defines up to $s = \prod_{i=1}^{N} l_i$ distinct strata, some of which may be structurally empty if certain combinations of factor levels are impossible.

2.1.4 | Prognostic balance, accidental bias, and statistical efficiency

Prognostic balance is achieved when the observed distribution of prognostic factors is similar in all randomized treatment groups. Prognostic balance is often considered a hallmark of a well-conducted trial. Yet, prognostic imbalances could occur through the play of chance, in a trial using completely random allocation, without raising any statistical issues

or implying any flaw in trial design and execution. Still, clinical trials with substantial imbalances can be the subject of criticism. An extensive list of controversial clinical trials can be found in Berger.[31]

Major prognostic imbalances, however, are unlikely to occur just by chance, and the treatment-allocation procedures discussed in this paper all achieve prognostic balance asymptotically. In practice, balance may be desired with limited sample sizes, either in small trials (eg, randomized phase 2 trials in oncology, cross-over trials in chronic diseases, or randomized trials in orphan indications), or at the time of interim analyses in large trials. Note that no allocation procedure will systematically reach exact balance as is shown further in Section 3.4.1 (Table 3) and Section 3.6.3 (Table 6), for simulated trials of size 500, and for 50 actual clinical trials with sample sizes ranging from <50 to >500 patients.

A valid reason to seek balance in a randomized trial is to avoid accidental bias in the estimation of the treatment effect due to the confounding effects of prognostic factors. Accidental bias affects the estimate of the treatment effect if all the following conditions apply: (1) the distribution of some baseline factors differs across randomized treatment groups, (2) the factors that are imbalanced have a prognostic impact on the outcome of interest, and (3) the method used to estimate the treatment effect does not adjust for these factors. Conditions (2) and (3) imply that accidental bias can be controlled by adjusting the analyses for important prognostic factors. Hence, if it were for accidental bias alone, there would be no need to seek balance between the randomized treatment groups, providing the analysis was adjusted for factors of prognostic importance. However, such adjusted analyses are less straightforward than simpler, unadjusted analyses. They may be suspicious if they are data-driven and may not be possible in small trials. Reducing the probability of accidental bias may therefore be a better strategy than adjusting for it. In contrast to other experiments that can be repeated if need be, clinical trials are seldom repeated identically, for obvious ethical reasons. However, when selecting a randomization procedure to achieve prognostic balance, caution should be taken to avoid increasing the predictability of allocations and potentially introducing selection bias (see Section 2.1.5).

Prognostic balance may be sought to avoid accidental bias, but does it also increase the power of the test for a treatment effect? In nonlinear models or generalized linear models commonly used in clinical trials (eg, logistic regression and survival models), balance does not minimize the variance of the estimate of the treatment effect (see section 3.2 in Robertson et al[33]). In such cases, optimal treatment-allocation procedures have been proposed to maximize statistical efficiency.[34-37] These procedures outperform other randomization procedures, but the impact of these procedures on power is negligible in practice for the commonly observed moderate and small treatment effects of interest.[7,38] We will confirm the negligible impact of balancing on power through simulations in Section 3 Note that for more extreme treatment effects or extreme heterogeneity in the variances, imbalanced and balanced allocations can have a larger impact on the power of the test.[38] Hence, if it were for power alone, there would be no need to seek balance between the randomized treatment groups for most trials—if anything, one could even consider creating imbalances to maximize efficiency when nonlinear models are used. Note that the trade-off between increasing power and balancing treatments is a general consideration in prioritizing the objectives of any treatment-allocation procedure.[39,40]

2.1.5 | Predictability and selection bias

Selection bias can affect the estimate of the treatment effect if the following conditions apply: (1) the physicians who participate in the trial can select the patients they decide to enter, for instance based on some important prognostic characteristic, and (2) they can do so with foreknowledge of the treatment allocations. Note that selection bias is intentional while accidental bias is due to chance alone. The choice of a treatment-allocation procedure has a major bearing on selection bias because some procedures lead to predictable allocations.[41] For instance, in a trial comparing two treatments randomly allocated in blocks of 4 treatments, the last treatment in a block is always predictable, while the penultimate treatment in a block is predictable in one third of the blocks. Hence, on average, such a design leads to 1 out of every 3 treatments being predictable. For this reason, using stratified blocked randomization is not recommended in open-label trials when the participating physicians have access to the history of previous treatment allocations. For the same reason, the size of the blocks is usually kept masked to participating physicians or is varied randomly over the course of the trial.

2.1.6 | Allocation ratio

The allocation ratio reflects the desired proportion of patients in each treatment group. More often, equal allocation (1:1) is used when there are two treatment groups, with allocation of patients to either group with equal probability.

However, unequal allocation ratios, such as 2:1 or 3:1, can be used, especially in early-phase trials, where it may be valuable to allocate more patients to an unknown experimental treatment than to a well-known standard of care. The limited availability of an experimental treatment might also justify unequal allocation ratios favoring the allocation of patients to the control arm (eg, 1:2).

2.2 | Treatment-allocation procedures

2.2.1 | Objectives

Treatment-allocation procedures may purport to achieve several objectives, and the choice of a procedure may depend on which of these objectives is most important in a particular context:

- include a random element in the allocation in such a way that no allocation is predictable, thereby decreasing the risk of selection bias
- achieve the target allocation ratio, for instance equal numbers of patients in the two randomized groups of a trial with 1:1 allocation ratio
- achieve prognostic balance between the randomized groups, thereby minimizing the risk of accidental bias
- provide an inferential basis for the statistical tests used to compare the randomized groups
- maximize the efficiency of these tests

All these objectives can be achieved to some extent, but a focus on one of these objectives may imply compromising some of the other objectives.[33,40]

2.2.2 | Common allocation procedures

The simplest treatment-allocation procedure is completely random allocation, which can easily accommodate unequal allocation ratios. The major advantage of completely random allocation is that there can be no foreknowledge of treatment allocations, thus ensuring full concealment. This feature is considered by some as an overriding consideration when choosing a treatment allocation procedure (see Section 2.1.5).[31] Completely random allocation protects against selection bias but it can result, with non-negligible probability, in deviations from the desired allocation ratio and in covariate imbalances, especially in small samples, potentially leading to accidental bias.[21,27]

All other treatment-allocation procedures are "restricted" in some sense. Blocked randomization, the most common form of restricted randomization, is used to ensure close adherence to the desired ratio of patients in each treatment group at any time during the trial. When selected prognostic factors need to be accounted for, stratified blocked randomization is used to ensure good balance of prognostic factors between treatment groups, reducing the probability of accidental bias and leading to balanced allocation within strata.[22] Although this procedure is in very common use, it can be unsatisfactory with regard to predictability of treatment allocations (see Section 2.1.5). This technique is useful when very few factors of major prognostic importance must be accounted for,[42] but it does not ensure good balance when the number of strata defined by the cross-classification of prognostic factors becomes too large.[43]

Minimization, described independently by Taves[44] and by Pocock and Simon,[45] is a dynamic treatment-allocation procedure that minimizes the overall imbalance on selected baseline factor margins between treatment groups. The procedure can be deterministic or stochastic. Although the Consolidated Standards of Reporting Trials (CONSORT) state that "*trials that use minimization are considered methodologically equivalent to randomized trials, even when a random element is not incorporated*",[46] it is considered good practice to use stochastic minimization to avoid any concerns that deterministic allocations may have induced selection bias. The Food and Drug Administration (FDA) guidance on adaptive trials (2019) states that "*predictability (in a covariate-adaptive algorithm) can be mitigated with an additional random component to prevent perfectly deterministic treatment assignment*".[47] Therefore, this tutorial will focus only on stochastic minimization. A formal exposition of minimization was given by Pocock and Simon,[45] and a simplified algorithm was proposed by Freedman and White.[48,49] Several extensions of minimization have been proposed to differentiate the baseline factors according to their relative prognostic importance (see Section 5).[50,51] Another set of covariate-adaptive

procedures is based on an 'imbalance score', defined as a weighted average of imbalances within factor margins, strata, and the study.[26,30] Minimization in this tutorial refers to the method proposed by Pocock and Simon that minimizes the average of marginal imbalances.[45]

2.2.3 | Implementation of the different treatment-allocation algorithms

Completely random allocation

In this case, one of m treatments is randomly chosen with probability $1/m$ for an equal allocation ratio or with the appropriate probability for an unequal ratio. A prespecified list with this random order of treatment allocations can be generated, and a patient presenting for randomization will be assigned to the next treatment on the list. Today, lists of treatment allocations are no longer prepared ahead of time, and patients are randomized on-line in real time, but the algorithm for treatment allocation is essentially the same.

Stratified blocked designs

In stratified permuted-block randomization, trial participants are subdivided into strata, and then randomization is further restricted by the blocks so that treatment assignments are balanced within blocks and over time. For a trial with t treatment arms, each block has a size of $t \times b$ containing b allocations to each arm. The orders of treatments within each block are randomly permuted. One major choice to be made is the block size. If $b = 1$, strata are balanced after every multiple of t allocations. The last ($t \times b$th) allocation in a block is always deterministic, and in some blocks earlier treatment allocations are also deterministic. Larger block sizes increase allocation randomness but decrease the treatment balance, as at the end of recruitment a larger proportion of the block can remain unfilled.

Consider the example of a trial with two treatments, A and B, with age (≤ 65/>65) and sex (F/M) as important prognostic factors used for stratifying treatment allocation. This design has 4 strata and within each stratum, permuted blocks of size $2 \times b$ are randomly generated prior to opening recruitment. For instance, if $b = 4$:

Stratum 1 (F, ≤ 65): | AABABBBA | ABBABABA | BABABABA | ...
Stratum 2 (F, >65): | BBAABBAA | BAABABAB | ABABABAB | ...
Stratum 3 (M, ≤ 65): | ABBBAABA | ABAABBAB | BAABBAAB | ...
Stratum 4 (M, >65): | BBBABAAA | ABABAABB | AABBAABB | ...

When a patient is eligible for randomization, the next available treatment in their stratum is assigned. For example, female patients 65 or younger (stratum 1) will be allocated treatments A, A, B, A etc. For unequal allocation ratios, each block would contain correspondingly unequal numbers of the available treatments.

Minimization

This section describes commonly used implementations of minimization, for those unfamiliar with the procedure. More elaborate descriptions of minimization are provided, for example, in Han et al.[52] Important prognostic factors are identified before the trial starts, and assignment of a new patient to a treatment group is determined to minimize the differences between the groups in terms of these factors, by balancing allocation over factor margins. Unlike stratified randomization, minimization aims to minimize the total imbalance for all factors simultaneously instead of considering mutually exclusive strata. Pocock and Simon[45] defined a more general procedure, but the most commonly used approaches for implementation are the range and variance methods, introduced with an illustrative example below.

Consider again the example of a trial with two treatments, A and B, with age (≤ 65/>65) and sex (F/M) as important prognostic factors. Center is added as a factor for which balanced treatment allocation is desirable. A 60-year-old woman in center XYZ is ready to be randomized into the trial that has the following status (Table 1):

The goal of minimization is to minimize the total imbalance on some scale. The range method minimizes the sum of the absolute values of the imbalances ($\delta_i, i = A, B$):

- If treatment A is allocated, $\delta_A = |24 - 22| + |56 - 54| + |17 - 20| = 2 + 2 + 3 = 7$;
- If treatment B is allocated, $\delta_B = |23 - 23| + |55 - 55| + |16 - 21| = 0 + 0 + 5 = 5$;
- Since $\delta_B < \delta_A$, allocation of treatment B is preferred with the range method.

TABLE 1 Illustrative example of the range and variance implementation of minimization, showing current status of a two-arm trial and imbalances if next patient is assigned to arm A or arm B.

Number of patients already allocated to	A	B	Imbalance if next patient allocated to A	Imbalance if next patient allocated to B
Age: ≤65	23	22	2 (|24–22|)	0 (|23–23|)
Gender: Female	55	54	2 (|56–54|)	0 (|55–55|)
Center: XYZ	16	20	3 (|17–20|)	5 (|16–21|)
	94 (T_A)	96 (T_B)	7 (δ_A)	5 (δ_B)

Note: T_A/T_B: sum of number of patients assigned to treatment A/B corresponding to factor levels of the next patient. Note that T_A/T_B are not the total number of patients randomized to A/B due to overlap between the rows.

The variance method minimizes the sum of the squares of the imbalances ($\sigma_i^2, i = A, B$):

- If treatment A is allocated, $\sigma_A^2 = (24 - 22)^2 + (56 - 54)^2 + (17 - 20)^2 = 4 + 4 + 9 = 17$;
- If treatment B is allocated,$\sigma_B^2 = (23 - 23)^2 + (55 - 55)^2 + (16 - 21)^2 = 0 + 0 + 25 = 25$;
- Since $\sigma_A^2 < \sigma_B^2$, allocation of treatment A is preferred with the variance method.

Of note, the different treatment preferences between the two methods are not an inherent feature of the method, but rather a topical occurrence. In general, simulation studies have shown similar results for the range and variance methods, perhaps with a slightly better performance for the variance method.[25,34] If the variance method is used, the preferred treatment can be obtained through a simple summation of the numbers under each treatment already allocated to A and B in each factor.[49] Denote these total allocations by T_A and T_B, respectively (note that these sums are not numbers of patients, given the overlap between table rows). In the above example (Table 1), $T_A = 94$ and $T_B = 96$, and since $T_A < T_B$, allocation of treatment A is preferred. Note that if $\delta_A = \delta_B$ in the range method, or $\sigma_A^2 = \sigma_B^2$ in the variance method, no treatment is preferred over the other.

The next step consists in allocating the preferred treatment with probability p (with $0.5 < p \leq 1$) and the other treatment with probability 1–p. In case there is no preferred arm, allocate A or B at random with probability 0.5.

- If $p = 1$, the allocation is deterministic (as in Taves[44]) and always goes to the preferred treatment;
- If $0.5 < p < 1$, a stochastic rather than a deterministic implementation of minimization is obtained.

The algorithms (range and variance methods) can easily be extended to the following situations:

- More than two treatment groups: the preferred treatment is the treatment with the lowest total allocations T_i;
- Different weights for different stratification factors: calculate total imbalances as weighted (instead of unweighted) sums across factors;
- Unequal allocation ratios: create as many pseudo-treatment groups as needed to obtain the desired allocation ratio, and then combine identical treatment groups (see Section 5.2).

More formally, the variance method for a study with equal allocation to two treatment groups can be described as follows. At the time of entering a new patient in the trial, let N_{ij} denote the number of patients already assigned to treatment j ($j = A, B$) with an identical value for prognostic factor i ($i = 1, \ldots, F$) as the new patient, and let $T_j = \sum_{i=1}^{F} N_{ij}$. For this patient, if $T_A < T_B$, allocate treatment A with probability $0.5 + \beta$ and treatment B with probability $0.5 - \beta$; if $T_A > T_B$, allocate treatment B with probability $0.5 + \beta$ and treatment A with probability $0.5 - \beta$; and if $T_A = T_B$ allocate treatment A or B with probability 0.5. Note that this algorithm essentially uses a biased-coin approach, with β ($0 \leq \beta \leq 0.5$) denoting the bias,[32,53] with the following possible choices for β in this algorithm:

- $\beta = 0$: equal to completely random allocation (both treatments selected with equal probability regardless of imbalances);
- $\beta = 0.5$: deterministic minimization;
- $0 < \beta < 0.5$: stochastic minimization (with commonly chosen values for β in the range of 0.3 to 0.4).

In case the study has more than two treatment groups, the biased-coin algorithm can be generalized as follows. At the time of entering a new patient in the trial, let N_{ij} denote the number of patients already assigned to treatment j ($j = A, \ldots, N$) with an identical value for prognostic factor i ($i = 1, \ldots, F$) as the new patient. Let $T_j = \sum_{i=1}^{F} N_{ij}$, $m = min\{T_j \, (j = A, \ldots, N)\}$ and $\mathcal{M}$ be the set of treatments j ($j = A, \ldots, N$) such that $T_j = m$, and $\mathcal{N}$ be the (possibly empty) set of other treatments. To allocate a treatment to the new patient, one of the two sets of treatments, $\mathcal{M}$ or $\mathcal{N}$ is first chosen: if $\mathcal{N}$ is empty, $\mathcal{M}$ is chosen; otherwise, $\mathcal{M}$ is chosen with probability $0.5 + \beta$ and $\mathcal{N}$ is chosen with probability $0.5 - \beta$. Thereafter, any of the treatments in the chosen set ($\mathcal{M}$ or $\mathcal{N}$) is selected at random with equal probability.

3 | OPERATING CHARACTERISTICS

3.1 | Case study: Design of a randomized clinical trial in oncology

We consider the choice of a treatment allocation procedure for a randomized trial comparing two chemotherapies for patients with advanced ovarian cancer. Data on such patients were available from a previously published meta-analysis (Ovarian Cancer Meta-Analysis Project).[54] Four trials were included in the meta-analysis, all comparing a control group receiving the two-drug combination of cyclophosphamide and cisplatin (CP) vs an experimental group receiving the three-drug combination of cyclophosphamide, doxorubicin and cisplatin (CAP). Data on the following baseline variables had been collected for every patient: patient identification, center, date of randomization, age, performance status, extent of residual disease after debulking surgery, histologic cell type, histologic cell differentiation, and International Federation of Gynecology and Obstetrics (FIGO) stage. Data were also available on the treatment assigned by randomization (which was ignored to regenerate allocations with the minimization algorithm under study) and the outcomes of interest: date of death or last visit and survival status.

Table 2 provides an overview of baseline characteristics for all 1198 patients available in the meta-analysis, as well as the χ2 score statistic for the prognostic impact of each characteristic (taken in isolation) on survival. The interest of this case study is that all the baseline variables had a highly significant prognostic impact on survival. A prognostic score was calculated for each patient by fitting a Cox (proportional hazards) regression model, starting from the null model with all baseline characteristics, and eliminating covariates from the model using a step-down procedure. Patients with missing values for the retained covariates were excluded from the estimation of the model parameters. All baseline variables were kept in the model: age (*Age*, $p < 0.0001$), performance status (*PS*, $p < 0.0001$), extent of residual disease (*RD*, $p < .0001$), histologic grade (*HG*, $p < 0.005$) and FIGO stage (*FIGO*, $p < 0.05$). A prognostic score was calculated as a linear combination of prognostic factors, with age in years and the levels of other factors coded from 1 (lowest level) to 3 (highest level):

$$score = 0.016 \times Age + 0.37 \times PS + 0.48 \times RD + 0.14 \times HG + 0.24 \times FIGO.$$

The final model was used to divide the patient population into quintiles defining distinct prognostic groups, respectively labeled "best", "good", "intermediate", "bad", and "worst" prognosis.

3.2 | Treatment-allocation procedures considered for this trial

We considered three treatment-allocation procedures when designing a trial of 500 patients entered in 20 centers to compare an experimental treatment E with a control treatment C. We assumed that the sample size in each center followed a skewed distribution, with a majority of centers entering few patients and a few centers entering the majority of patients as is typically the case in cancer clinical trials. Simple randomization consisted of allocating E or C at random to each patient, with probability 0.5. Stratified blocked randomization consisted of forming blocks of size 4 (2E, 2C) within strata arising from the cross-classification of center and selected prognostic variables from Table 2. We implemented minimization using the variance of imbalances,[48,49] (see Section 2.2.3 for details) with bias β set to 0.4 (stochastic minimization).

TABLE 2 Prognostic impact of baseline variables on survival, as assessed by a Cox model.

Variable	Ordered categories	Number of patients (percentage)	Cox model score statistic
Age (years)	< 55	525 (0.44)	53.95
	55−65	419 (0.35)	
	≥ 65	248 (0.21)	
	Missing	6 (<0.01)	
Extent of residual disease (cm)	None	191 (0.16)	101.44
	≤ 2	463 (0.39)	
	>2	523 (0.43)	
	Missing	21 (0.02)	
Histologic grade	Good	206 (0.17)	40.78
	Intermediate	420 (0.34)	
	Poor	463 (0.39)	
	Missing	127 (0.11)	
FIGO stage	III	996 (0.83)	38.77
	IV	157 (0.13)	
	I, II or missing	45 (0.04)	
ECOG Performance Status	0	770 (0.64)	32.60
	1	313 (0.26)	
	≥ 2	59 (0.05)	
	Missing	56 (0.05)	
Final prognostic score, including age, extent residual disease, histologic grade, FIGO stage and ECOG	Best	176 (0.15)	192.35
	Good	262 (0.22)	
	Intermediate	260 (0.22)	
	Bad	211 (0.17)	
	Worst	177 (0.15)	
	Missing	112 (0.09)	

Note: All variables are ordered categorical or continuous, and patients with missing values are excluded from the model. The score statistic has a χ^2 distribution with 1° of freedom.

3.3 | Methods

3.3.1 | Simulations

We generated 500 trials, each of them with 500 patients sampled at random, with replacement, from the meta-analysis database. We took the survival times equal to their values in the meta-analysis database for patients allocated to the control group and divided by the hazard ratio (*HR*) for patients allocated to the experimental group (survival times in the meta-analysis were found to closely follow an exponential distribution). We assumed the HR was equal to 1.0 under H_0 and equal to 0.7 under H_A. The meta-analysis database had a long follow-up (average follow-up more than 12 years). To make the simulations realistic, we censored survival times to account for a uniform entry of the patients over a 2-year period, and an additional 4-year follow-up period. In each simulated trial, we compared the treatment groups using a Wald test for the treatment coefficient in a Cox model, with and without stratification by baseline covariates as explained below. We assumed *HR* was constant, and the sample size of 500 patients was chosen to have a power greater than 0.8 of detecting *HR* ≤ 0.7, using a two-tailed significance level of 0.05.

3.3.2 | Measure of imbalance

We defined the imbalance between the randomized treatments as $\delta = |n_E - n_C|$, where n_E and n_C are the number of patients allocated to experimental and control, respectively. When looking at S subsets of patients defined either by prognostic-factor levels (eg, patients with the same baseline performance status, or in a given center) or by strata arising from the cross-classification of several factors (eg, patients with the same baseline performance status in a given center), the imbalance was defined as $\delta_S = \max_s \{|n_{Es} - n_{Cs}|\}$ for $s = 1, \ldots, S$.

3.3.3 | Type-I error probability and power of the statistical test

The proportion of simulated trials rejecting H_0 under H_0 gave the type-I error probability of the test for the treatment effect. The proportion of simulated trials rejecting H_0 under H_A gave the empirical power of the test for the treatment effect.

3.3.4 | Measure of predictability

We measured the predictability of the next treatment allocation by the probability of a correct guess for an investigator using an optimal guessing strategy.[55] Completely random allocation is not considered here, as there is no optimal guessing strategy if an allocation only depends on randomness. If the guesser only had access to the total number of patients already allocated by treatment group, the optimal guessing strategy would be:

- For both stratified block allocation and minimization, pick the treatment that minimizes overall imbalance (ie, difference between the total numbers of patients randomized to experimental and control), if there is such a treatment; otherwise, choose a treatment at random.

If the guesser had knowledge of both the allocation algorithm and the history of all previous allocations, the optimal guessing strategy would depend on the allocation algorithm:

- For stratified blocked allocation, pick the treatment that minimizes imbalance in the current block corresponding to the patient characteristics, if there is such a treatment; otherwise, choose a treatment at random;
- For minimization, pick the treatment that minimizes overall imbalance using the minimization algorithm, if there is such a treatment; otherwise, choose a treatment at random.

Note that in well-conducted multicenter trials, it should be impossible for any investigator to have all knowledge needed for the optimal guessing strategy, even when the study is open-label. In this section, predictability is rigorously defined to quantify its potential impact. However, in practice it does not matter whether the next allocation is predictable but rather whether those in a position to influence recruitment believe they know what will happen next with higher chance, regardless of this being correct or not. Acting upon a belief that is not correct could introduce bias nonetheless.

3.4 | Results

3.4.1 | Balance

Table 3 shows the mean imbalance (number of patients) overall, within centers (see Section 2.1.2), within factor levels (see Section 2.1.1) and within strata (see Section 2.1.3), for the various treatment allocation procedures. Minimization achieves similar or better balance than stratified blocked allocation, except within strata, where the difference between the two procedures is trivial. These results are in agreement with those from other publications using simulations.[56-58]

TABLE 3 Mean imbalances (number of patients) overall, within centers, within factor levels and within strata in a trial of 500 patients, based on 500 simulations.

	Overall			Within centers			Within factor levels			Within strata		
Factors used	CRA	SBA	MIN	CRA	SBA	MIN	CRA	SBA	MIN	CRA	SBA	MIN
0	15.4			11.6			23.1			4.0		
Center		3.1	4.0		1.8	2.6		17.1	17.6		3.9	3.9
+ score		6.9	2.2		4.2	3.2		17.6	15.8		3.7	3.9
+ 1 factor		5.4	1.9		3.2	4.1		16.8	15.3		3.8	3.9
+ 3 factors		10.5	1.4		7.0	4.8		17.3	11.8		3.2	3.8
+ 5 factors		13.6	1.3		9.5	3.2		20.0	4.1		2.0	3.8

Abbreviations: CRA, completely random allocation; MIN, minimization; SBA, stratified blocked allocation.

TABLE 4 Type-I error probability and power of the statistical test for a treatment difference.

	Type-I error probability						Power					
	Unstratified test			Stratified test*			Unstratified test			Stratified test*		
Factors used	CRA	SBA	MIN	CRA	SBA	MIN	CRA	SBA	MIN	CRA	SBA	MIN
None	0.052			0.043			0.72			0.79		
Center		0.051	0.049		0.046	0.040		0.74	0.73		0.79	0.79
+ score		0.049	0.026		0.047	0.040		0.74	0.73		0.77	0.79
+ 1 factor		0.046	0.044		0.035	0.042		0.75	0.75		0.81	0.80
+ 3 factors		0.042	0.036		0.034	0.036		0.74	0.75		0.80	0.82
+ 5 factors		0.041	0.035		0.034	0.035		0.76	0.75		0.81	0.81

Abbreviations: CRA, completely random allocation; MIN, minimization; SBA, stratified blocked allocation.

*Test stratified for 1 factor.

3.4.2 | Type-I error probability and power

Table 4 shows the type-I error probability and power of the statistical test for a treatment difference for the various treatment-allocation procedures, using either an unstratified Cox model or a Cox model stratified by one factor. As compared with completely random allocation, the type-I error probability of the test is smaller for stratified blocked allocation and for minimization, due to the balance achieved by these procedures, which reduces the likelihood of accidental bias leading to false-positive results. For unstratified tests, the type-I error probability decreases with increasing number of factors used in the randomization procedure. The power of the test is largely unaffected by the treatment-allocation procedure but is increased if the test is stratified for one factor, regardless of the treatment-allocation procedure. This observation is consistent with the well-known fact that a stratified analysis gains power due to a reduction in the variance of the treatment effect.

In this example, an analysis stratified by all five factors would increase the power even further, again with no major difference between the various treatment-allocation procedures (0.86 for completely random allocation, 0.85 for stratified blocked allocation, and 0.89 for minimization, results not shown in Table 4). Over-stratification of the analysis, however, induces a loss of power due to the large number of strata with few patients. In this example, an analysis stratified by center, score, and all five factors (obviously an extreme over-stratification), would seriously decrease the power (0.41 for completely random allocation, 0.46 for stratified blocked allocation, and 0.45 for minimization, results not shown in Table 4).

3.4.3 | Predictability

Table 5 shows the probability of correctly guessing the next treatment allocation. This probability is equal to 0.5 by definition for completely random allocation. Both stratified blocked allocation and minimization increase predictability,

TABLE 5 Probability of correctly guessing the next treatment allocation.

Factors used	No information available by treatment group			Total numbers of patients known by treatment group			Full knowledge of allocation algorithm and history		
	CRA	SBA	MIN	CRA	SBA	MIN	CRA	SBA	MIN
None	0.500			0.500			0.500		
1		0.500	0.500		0.624	0.631		0.707	0.722
3		0.500	0.500		0.542	0.647		0.699	0.832
5		0.500	0.500		0.521	0.648		0.672	0.858

Abbreviations: CRA, completely random allocation; MIN, minimization; SBA, stratified blocked allocation.

especially when the investigator has full knowledge of the allocation algorithm and of the history of previous allocations. For stratified blocked allocation, predictability decreases when the number of factors used in the allocation increases; for minimization, it is the opposite. This is due to the better balancing properties of minimization (Section 3.4.1). Minimization has been criticized because it can increase predictability.[9] The same criticism holds for stratified permuted blocks.[59] However, as shown in Table 5, predictability is only an issue when minimization is used in the uncommon scenario where investigators participating in a trial have (1) knowledge of the treatment allocation algorithm, and (2) information by treatment group on patients allocated so far.[60] If predictability is a concern for a specific trial, modified algorithms are available to reduce predictability.[61,62]

3.5 | Interpretation

Section 3.4.2 shows the benefit of achieving prognostic balance on the type-I error probability of the statistical test used to compare treatment groups and the lack of effect of such balancing on the power of the test. Section 3.4.3 shows, however, that using procedures that result in such balance comes at the cost of unacceptably high levels of predictability if investigators have full or even partial information on the history of previous allocations, including a simple count of the number of patients randomized by treatment group. All such information, as well as details of the treatment allocation algorithm, must therefore be concealed from investigators to prevent selection bias, particularly in open-label and in single-center trials. In properly blinded multicenter trials, there is no reason to suspect that the level of predictability is materially higher than for completely random allocation, though no procedure guarantees unpredictability of treatment allocations as effectively as completely random allocation. Is the impact of balancing on the type-I error probability of the statistical test sufficient to warrant the use of stratified blocked randomization or minimization? It seems debatable, though tight control of the type-I error is an overriding consideration in the approval of a new treatment. In addition, regulatory and health-technology assessment (HTA) agencies may be more easily assured of scientific validity when treatment allocation is balanced for important prognostic factors and across geographic regions.

Balancing has negligible effect on power, so it may not be worth the trouble from the point of view of a trial sponsor. In contrast, stratification of the statistical tests may have a large effect on power, at least when baseline factors have a substantial prognostic impact on the outcome of interest, as in advanced ovarian cancer. As expected, stratification of the analysis increased power equally for all treatment-allocation procedures. It is often claimed that the factors used for treatment allocation must be included in the statistical model used to compare treatment groups. Although this is a sound strategy, when possible, it does not follow that it should always be used uncritically, since balancing treatment groups for prognostic factors aims at reducing accidental bias, while adjusting the analysis for prognostic factors aims at increasing the statistical power of the test. These two objectives may be best met for different numbers of prognostic factors in a given study.

The type-I error probability was controlled in all situations, regardless of the treatment-allocation algorithm, number of factors used and stratification of the (asymptotic) test. As such, our results confirm that minimization and permuted blocks control the type-I error probability of asymptotic tests.[63,64] Of note, permuted blocks and minimization show type-I error probabilities below the nominal 0.05 level, illustrating that asymptotic tests can be conservative in these cases. Research on statistical inference under covariate-adaptive randomization has shown that asymptotic tests not stratified for

all balancing factors are usually conservative, albeit less so than under complete randomization.[11-13] The conservativeness of the tests is due to the reduced variance of the treatment effect, with the achieved marginal balance across treatment groups, enhancing the comparability of different treatment groups. When an unstratified linear model is implemented for inference, the covariate-adaptive randomization scheme is ignored and an inflated model-based variance estimate is used, resulting in a conservative control of the type-I error probability. Stratifying the analysis for all balancing factors restores the type-I error probability to the nominal level with an asymptotically normally distributed test statistic.[14,65] A bootstrap- based test and Lasso regression have been suggested to further improve the efficiency of asymptotic tests following a stratified randomization.

To the best of our knowledge, these approaches have not been used in clinical trials. Randomization tests provide a simple and robust testing approach and, as such, they can advantageously be used instead of asymptotic tests.

3.6 | Experience with minimization

To assess the performance of minimization in practice, we reviewed trials in which we used minimization for treatment allocation. We looked at all such trials randomized at our institution (IDDI) between August 28, 2000 and March first, 2021 ($N = 61$). Eleven trials were excluded for different reasons: trials that stopped prematurely ($N = 2$), trials randomized without any stratification factor ($N = 2$), and trials with less than 25 patients ($N = 5$). Hence, 50 trials using minimization for treatment allocation with a total of 20 469 patients could be used for this analysis. The sample sizes of the studies ranged from 28 to 3509 patients, with a mean and median sample size of 409.4 and 155, respectively. All studies were multicenter trials, with the number of centers ranging from 3 to 567. Most trials ($N = 35$) were two-arm trials, with 13 and 2 studies having three and four arms, respectively. Forty-one trials had an equal allocation ratio, and nine had unequal ratios (2:1 in six cases, and one each 3:1, 4:1, and 2:2:1). Among the indications treated in the 50 trials analyzed, oncology and ophthalmology were the most frequent.

3.6.1 | Treatment-allocation procedure used in these trials

All minimization algorithms were stochastic and were implemented using the variance method. The number of factors used in the algorithms ranged from 1 to 6, with a mean and median number of factors of 3.0 and 3, respectively. The total number of strata per trial varied from 2 to 3402, with a mean and median of 315.8 and 64, respectively. The large number of strata is caused by the inclusion of center as stratification factor in 41 of the 50 trials. Without the factor center, the number of strata per trial ranges from 0 to 48, with a median and mean of 6.0 and 4, respectively. A total of 13 studies had the one-level factor "study" included, to improve the overall balance of the treatment allocations.

3.6.2 | Methods

We defined the overall imbalance between the randomized treatments arms as $\delta = |n_E \times R_C - n_C \times R_E|$, with n_E and n_C the number of subjects allocated to experimental and control groups, respectively, and where R_E and R_C are the integers in the randomization ratio (eg, for a ratio 2:1 favoring experimental treatment, $R_E = 2$ and $R_C = 1$). For studies with more than two treatments, we computed the imbalance for every pair of treatments and kept the maximum value. When looking at S subsets of patients defined by factor levels (eg, patients with the same baseline performance status), we defined the imbalance within factor levels as $\delta_S = \max_s \{|n_{Es} \times R_C - n_{Cs} \times R_E|\}$ for $s = 1, \ldots, S$. If a study contained more than two treatments, we computed the imbalance within factor levels for every pair of treatments and kept the maximum value. Finally, we also looked at the imbalance for a factor that was not included in the minimization algorithm. For 42 trials, patient age at enrolment was available and was not included in the minimization algorithm. For each of those studies, we calculated median age and divided patients in two subsets ($\leq$median or $>$ median), calculating the imbalance considering these two categories: $\delta_{age} = \max_s \{|n_{Es} \times R_C - n_{Cs} \times R_E|\}$ for $s = 1, 2$. Again, if a study contained more than two treatments, we kept the maximum imbalance value.

3.6.3 | Results

Table 6 provides the overall imbalance and the imbalance within minimization factor levels and within age groups. The overall imbalance, or the deviation from the prespecified allocation ratio, ranged from one to six patients, with the highest mean imbalance for the largest trials (sample sizes above 500 patients). The maximum within-factor level imbalance ranged from one to seven patients, with a trend for higher imbalances for larger trials and trials including more factors and having more factor levels. No clear difference could be detected in imbalances between studies including or not the center as factor or between trials using equal and unequal allocation ratios. The imbalance for age groups, a factor not included in the treatment allocation, were higher than for factors included in the minimization algorithm, ranging from one to 21 patients, with a mean imbalance of 5.7 patients.

3.6.4 | Interpretation

Across the 50 trials, minimization appears to have achieved its objective of treatment balance within factor levels while preserving the overall treatment-allocation ratio. This holds true for the smallest trials, with sample sizes

TABLE 6 Mean imbalance in number of patients overall, maximum within factor level and maximum within age categories for 50 actual trials using minimization.

Category (Nb trials)	Overall imbalance		Within factor level imbalance		Age group imbalance		
	Range	Mean	Range	Mean	(Nb trials)[a]	Range	Mean
Overall (50)	1–6	1.8	1–7	3.1	(42)	1–21	5.7
By Sample Size (SS)							
SS < =50 (6)	0–3	1.3	1–4	1.8	(6)	1–6	2.2
50 < SS < =100 (12)	1–5	2.1	1–5	2.7	(12)	3–9	5.3
100 < SS < =200 (12)	0–4	1.2	1–4	2.2	(9)	1–8	5.2
200 < SS < =500 (12)	0–4	1.3	2–6	3.6	(10)	1–13	6.3
SS > 500 (8)	0–6	3.6	4–7	5.3	(5)	3–21	10.8
By nb factors							
1 factor (2)	2–2	2.0	1–2	1.5	(2)	4–7	5.5
2 factors (19)	0–5	2.1	1–5	2.7	(15)	1–13	5.2
3 factors (12)	0–5	2.2	1–7	3.6	(12)	1–21	7.0
>=4 factors (17)	0-6	1.3	1–6	3.4	(13)	1–10	5.2
By nb factor levels							
Nb levels <=10 (10)	0–2	1.3	1–2	1.5	(10)	1–8	3.8
10 < Nb levels <=50 (10)	0–6	2.6	1–6	2.6	(9)	1–13	6.8
50 < Nb levels <=300 (16)	0–4	1.3	2–4	3.0	(12)	1–11	5.0
Nb levels >300 (14)	0–5	2.3	3–7	4.6	(11)	1–21	7.5
Center included as factor							
Yes (41)	0–5	1.8	1–7	3.3	(34)	1–21	6.2
No (9)	1–6	1.9	1–6	2.0	(8)	1–8	3.9
By allocation ratio							
Equal (41)	0–5	1.8	1–7	3.1	(34)	1–21	6.4
Unequal (9)	0–6	2.1	1–6	2.8	(8)	1–8	3.0

[a] Imbalance for age group was calculated in 42 of 50 studies for which patient's age at enrolment was available and was not included in the minimization algorithm.

of up to 50 patients, as well as for the largest trials. The number of minimization factors and factor levels did not affect the achieved balance. These empirical results confirm earlier findings that properly implemented stochastic minimization algorithms result in small imbalances with respect to the variables used in the allocation process even in trials of small size and/or when the number of prognostic factors is large.[1,25,57,66] The inclusion of study center as factor, with hundreds of different centers, did not affect the achieved balance. These results further demonstrate that achieved balances are similar for equal and unequal allocation ratios, as expected when using the correct implementation with "allocation-ratio-preserving, biased-coin minimization" approach for unequal ratios (see Section 5.2).

As expected, for 42 studies in which age was not included as a factor in the minimization algorithm, the imbalances were larger than those found for factors included in the allocation algorithm, showing that including a factor in the allocation algorithm is worthwhile if balance is aimed for. For trials with a sample size above 50, the mean imbalance for age is higher compared with factors included in the algorithm and increases with the sample size. Note that only potentially prognostic factors are included in the allocation algorithm, hence it is inevitable that some baseline factors will not be approximately balanced across treatment groups. On the other hand, the mean imbalance for age remained, with a mean of 5.7 patients overall and 10.8 patients for the largest trials (sample size >500), well below the mean imbalance of 23.1 patients reported for the within-factor imbalance in simulated trials of 500 patients using completely random allocation (Section 3.2, Table 3). The lower imbalance found for age (Table 6), in comparison with imbalance in factor levels achieved by completely random allocation (Table 3), could be induced by a correlation between age and included prognostic factors (eg, performance status in oncology trials, visual acuity at baseline in ophthalmology trials), so that balancing for these prognostic factors also improved balance for age on average. This result concurs with the reported benefits of using covariate adaptive randomization or stratified blocked designs in terms of balancing unobserved covariates.[67]

4 | RANDOMIZATION TESTS

4.1 | Implementation and hypotheses

In a first step, the observed test statistic T is computed, for example, a log-rank test statistic for time-to-event endpoints (as used in the ovarian cancer trial in Section 3) or a Cochran-Mantel-Haenszel test statistic for binary endpoints (as used in the ophthalmology trial in this section). In a second step, the distribution of the test statistic under the null hypothesis is obtained by randomizing treatment assignment according to the original randomization protocol while keeping outcomes and covariates as observed and rederiving the test statistic for each randomization. For a one-sided test, the empirical P-value for the test statistic based on the randomization distribution is then S/R, where S denotes the number of replicates for which the test statistic obtained in the randomization procedure in Step 2 is equal to, or greater in absolute value than, the observed value of T obtained in Step 1, and R denotes the total number of replicates.

For a randomization test, the patients' order of entry is kept fixed in the simulations and the randomization procedure is reapplied, whereas for a standard permutation test, the treatment allocations are merely reshuffled. A permutation test hence assumes all possible treatment allocation sequences at the end of the trial are equiprobable while a randomization test relies on the act of randomization to be the basis of inference.[1,69] See the reference book by Rosenberg and Lachin[1] for a theoretical explanation and Wang, Rosenberger and Uschner (see section 6.4 in Reference 68) for an example on the difference between randomization and permutation tests. Although stratified blocked designs (like any other design) can also be analyzed by means of randomization tests, in practice this is seldom done.

Under the null hypothesis of no treatment effect, the randomization model states that patient responses are unaffected by any treatment received. In contrast to population-based tests, the hypothesis for a randomization test involves no parameters and is, in essence, a statement that the treatment assignments are independent of the patient outcomes.[1] The randomization distribution of a test statistic is always correct for the set of observations and as such, the type-I error probability is preserved as long as the study is not biased.[68] In this sense, the randomization test is considered "robust", whereas the validity of population-based tests is conditional on distributional assumptions.

4.2 | Case study: Analysis of a randomized clinical trial in ophthalmology

We consider a phase 3 trial in which patients with AMD were randomized between a control group receiving no treatment or one of three doses (denoted low, medium, high) of pegaptanib sodium, an anti-vascular endothelial growth factor pegylated aptamer. This treatment was administered through intra-ocular injections every 6 weeks. In order to preserve treatment masking, patients in the control group were submitted to a "sham" injection (with pressure applied to the eye but no intra-ocular injection) every 6 weeks. Visual acuity was the main outcome of interest, and the primary endpoint was the proportion of patients who had lost three lines (15 letters) of vision on a standard visual acuity chart after 1 year on trial. The full results of this trial, which recruited 586 patients at 58 sites between August 2001 and July 2002, have been reported.[70]

4.3 | Treatment-allocation procedure used in this trial

Patients were allocated to one of the four arms using stochastic minimization. Two disease-related factors were expected to have an impact on visual acuity and were therefore used as "minimization factors": lesion type (classic only, predominantly classic, or occult) and prior treatment with photocoagulation (yes or no). Centre was also added as a minimization factor. The minimization algorithm was implemented using the variance of imbalances as the metric of choice[48,49] and followed the generalization of the biased-coin algorithm for studies with two arms, introduced in Section 2.2.3, with the bias β set to 0.3.

4.4 | Methods

4.4.1 | Analysis of the actual trial with randomization and asymptotic tests

An asymptotic test assumes that the sample size is sufficiently large so that the test statistic converges to an appropriate limiting distribution. The asymptotic test used to compare proportions of patients was the Cochran-Mantel-Haenszel test, which follows a chi-square distribution with 1° of freedom under the null hypothesis. A total of three Cochran-Mantel-Haenszel tests were performed: unstratified, stratified for the two minimization factors, and stratified for the two minimization factors plus center. The three associated *P*-values were recorded and in practice, would be compared with the nominal type-I error probability, commonly set at a one-sided 0.025 significance level.

Randomization tests were carried out on 10 000 simulated trials in which the outcome and characteristics of all patients were kept unchanged alongside their order of entry, but their treatment was reallocated using the same stochastic minimization algorithm as in the actual trial.[69-71] Three Cochran-Mantel-Haenszel test statistics (unstratified and stratified for two and three factors) were calculated for each simulated trial, and the three sets of 10 000 values thus obtained provided an empirical distribution of the test statistics under the null hypothesis. The actual value of the test statistics was compared with this distribution to yield a randomization *P*-value. Note that the asymptotic tests are much less computationally intensive (only 3 tests performed based on the actual treatment allocation) compared with the randomization tests that require thousands of simulations to establish the empirical distribution of the test statistics under the null hypothesis (here, 10 000 simulated datasets with three tests statistics calculated for each dataset).

Note also that the *P*-values shown in Section 4 are not adjusted for multiplicity, although in clinical research a multiple-testing correction would be needed when testing three doses vs a control treatment.

4.4.2 | Comparison of randomization and asymptotic *P*-values in simulated trials

In order to further study the difference between *P*-values obtained with randomization and asymptotic tests, we simulated 500 studies by resampling from the original study data, stratified by treatment arm without replacement. This procedure was performed to sample datasets of 100 patients per treatment arm (study with 400 subjects) and smaller studies with 25 patients per treatment arm (study with 100 subjects). For each sampled dataset, the three treatment comparisons were

analyzed with an asymptotic Cochran-Mantel-Haenszel test and randomization test using the Cochran-Mantel-Haenszel test statistic. The analyses for datasets consisting of 400 patients were stratified for lesion type and prior photocoagulation therapy; the analyses of the smaller datasets of 100 patients were performed unstratified.

4.5 | Results

4.5.1 | Analysis of the actual trial with randomization tests

Table 7 displays the asymptotic P-values calculated with the Cochran-Mantel-Haenszel statistic stratified for lesion type and prior photocoagulation for the three pairwise comparisons of interest in the trial (each dose vs sham) and the randomization P-values for the same comparisons. The three doses differed in efficacy, with the low dose exhibiting the largest difference from control. For all three doses, the asymptotic P-values were quite close to the randomization P-values and would have led to the same conclusions (situation "A" in Table 7.

The same qualitative observations were made when the Cochran-Mantel-Haenszel statistic was stratified for center in addition to lesion type and prior photocoagulation, but as shown in Section 2 the stratification by center leads to a loss in power because of the large number of centers entering patients in this trial (situation "B" in Table 7). For unstratified tests (situation "C" in Table 7), the asymptotic P-values were again quite close to the randomization P-values and would have led to the same conclusions. Note that there was no clear difference in power between the unstratified test and the test stratified for the two minimization factors (situation "A"), which suggests an absent or modest prognostic effect of the minimization factors. Indeed, a chi-squared test showed no significant dependencies between minimization factors and vision loss (P-values of 0.3018 and 0.3153, for lesion type and prior photocoagulation treatment, respectively).

4.5.2 | Comparison of randomization and asymptotic P-values in simulated trials

Figure 1 displays scatterplots of the asymptotic vs the randomization P-values. The P-values for the three pairwise comparisons of each dose vs sham are plotted together, for a total of 1500 pairs of P-values. The top and bottom rows show results for the trials with 100 and 400 patients, respectively. For larger P-values, the relative difference between both types of tests is well within 10% for both sample sizes (dashed lines on Figure 1). For the lower P-values, such as those below 0.1, a few relative differences exceed 10%.

TABLE 7 Asymptotic and randomization P-values for different tests with treatment allocated using minimization in a clinical trial for patients with age-related macular degeneration.[71]

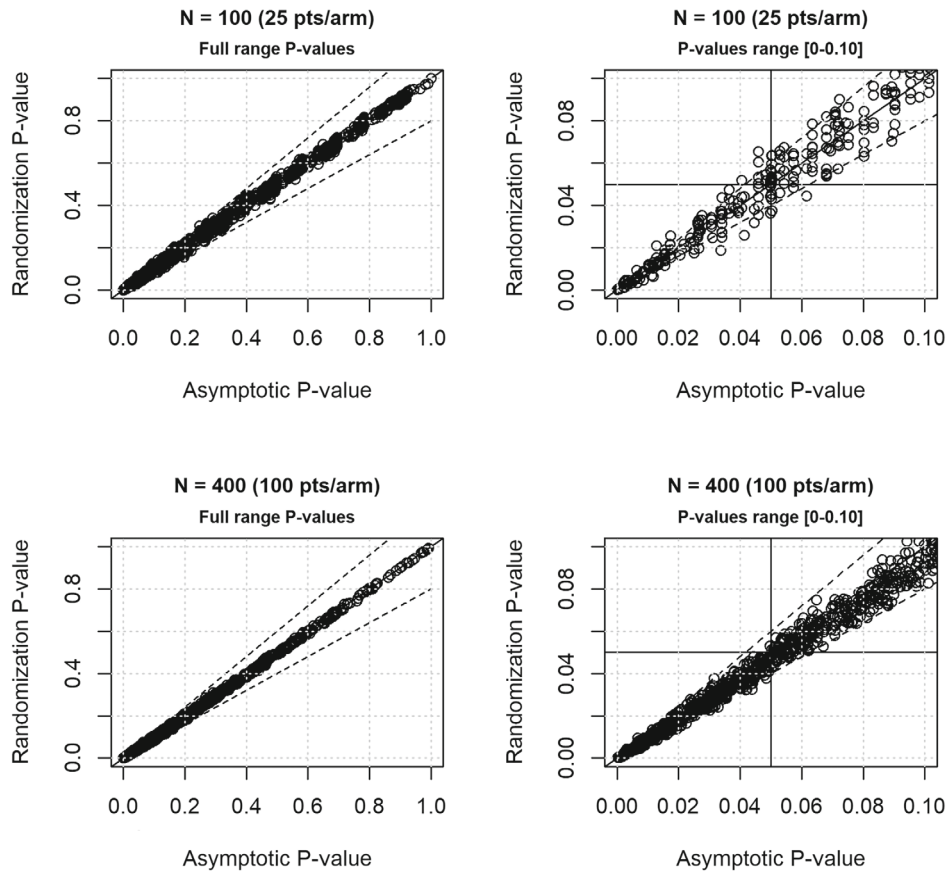| Situation	Pairwise Comparison	Asymptotic test	Randomization test
A: Test stratified for lesion type and prior photocoagulation*	Low vs sham	$P = 0.0156$	$P = 0.0130$
	Medium vs sham	$P = 0.0321$	$P = 0.0273$
	High vs sham	$P = 0.1292$	$P = 0.1174$
B: Test stratified for centre, lesion type and prior photocoagulation	Low vs sham	$P = 0.0317$	$P = 0.0279$
	Medium vs sham	$P = 0.0812$	$P = 0.0781$
	High vs sham	$P = 0.1672$	$P = 0.1619$
C: Unstratified test	Low vs sham	$P = 0.0148$	$P = 0.0127$
	Medium vs sham	$P = 0.0200$	$P = 0.0167$
	High vs sham	$P = 0.1391$	$P = 0.1320$

*Prespecified primary test.

FIGURE 1 Pairwise comparison of asymptotic and rerandomization *P*-values based on 500 simulated trials consisting of 4 arms and 3 comparisons each (1500 *P*-values). Top row: sample size of simulated trials with 25 patients per arm; bottom row: sample size of simulated trials with 100 patients per arm. Left column: full range (*P*-values [0, 1]); right column: zoomed to *P*-value range [0, 0.1]. Results for all three treatments comparisons shown together (3 × 500 or 1500 pairs of *P*-values). Dashed lines: 10% relative difference lines.

Table 8 summarizes the asymptotic and randomization *P*-values, as well as relative and absolute differences between them. The absolute difference was calculated as

$$Absolute\ Difference = Randomization\ p - Asymptotic\ p,$$

so that negative differences represent simulations for which the randomization *P*-value was lower. The relative difference was calculated as

$$Relative\ Difference = (Randomization\ p - Asymptotic\ p)/(Asymptotic\ p).$$

Median *P*-values and median differences are provided together with the interquartile range. Results are presented by asymptotic *P*-value size. Figure 2 depicts the distribution of absolute differences by asymptotic *P*-value size. Overall, the absolute differences are small, with an interquartile range (IQR) for studies of 100 patients within [−0.02, 0.02] and for the studies of 400 patients, within [−0.01, 0.01].

4.6 | Interpretation

In the analysis of the AMD trial data, randomization tests yielded *P*-values very close to the asymptotic *P*-values in all situations considered.

TABLE 8 Median and interquartile range (IQR) of asymptotic *P*-values, randomization test *P*-values, absolute and relative difference in *P*-values for 500 simulated trials with 4 treatment arms and 3 pairwise comparisons each (for a total of 1500 *P*-values).

Sample size (Total)	Pairwise Comparison	Asymptotic *P*-value median [IQR]	Randomization *P*-value median [IQR]	Relative difference in *P*-values median [IQR]	Absolute difference in *P*-values median [IQR]
100	All*	0.392 [0.167,0.671]	0.406 [0.170,0.671]	0.010 [−0.013,0.042]	0.003 [−0.004,0.013]
	P-value <0.05	0.025 [0.012,0.036]	0.022 [0.011,0.036]	−0.008 [−0.119,0.114]	0.000 [−0.002,0.003]
	0.05 ≤ *P*-value<0.1	0.072 [0.062,0.09]	0.074 [0.062,0.093]	0.052 [−0.055,0.131]	0.003 [−0.004,0.009]
	0.1 ≤ P-value<0.2	0.16 [0.124,0.170]	0.161 [0.129,0.182]	0.024 [−0.059,0.087]	0.004 [−0.008,0.013]
	P-value≥0.2	0.564 [0.363,0.780]	0.571 [0.375,0.791]	0.008 [−0.008,0.030]	0.005 [−0.004,0.016]
400	All*	0.097 [0.036,0.221]	0.090 [0.034,0.214]	−0.04 [−0.094, −0.008]	−0.004 [−0.008, −0.001]
	P-value <0.05	0.023 [0.012,0.035]	0.021 [0.010,0.032]	−0.095 [−0.163, −0.024]	−0.002 [−0.003,0.000]
	0.05 ≤ P-value<0.1	0.074 [0.061,0.085]	0.067 [0.057,0.079]	−0.068 [−0.113, −0.021]	−0.005 [−0.008, −0.001]
	0.1 ≤ P-value<0.2	0.140 [0.118,0.165]	0.134 [0.111,0.161]	−0.044 [−0.070, −0.013]	−0.006 [−0.010, −0.002]
	P-value≥0.2	0.338 [0.254,0.513]	0.332 [0.248,0.510]	−0.014 [−0.033, −0.002]	−0.005 [−0.010, −0.001]

Note: The trials were of size $N = 100$ or $N = 400$, and were simulated by random sampling from a clinical trial for patients with age-related macular degeneration.[71]

*1500 simulated *P*-values.

Other than for checking purposes, the choice of randomization tests can be advocated to reflect the actual treatment allocation procedure. If many factors are used in minimization, or if the number of levels of some factors is large (eg, centers), a test stratified for all factors may be much less efficient than a test stratified only for the most important prognostic factors. The same reasoning holds true for both asymptotic and randomization tests (compare situations "A" and "B" in Table 7 for both tests). It is therefore common practice, when minimization is used, to ignore some of the minimization factors for testing purposes, even if minimization allows for more factors in the actual allocation procedure. For instance, using a test stratified for center will not result in any gain in power unless there is a strong center effect and the number of centers in the trial is limited, which is an unusual (and arguably less desirable) situation. Note that in the AMD trial, the unstratified analysis had the smallest *P*-values across both test types showing that in case the stratification factors are not strongly prognostic for the outcome, stratifying a test can also result in a (slight) loss of efficiency.

Asymptotic and randomization tests could yield different results if there was a time trend in the outcome of interest, for instance if there was a shift in the characteristics of the patients entered over time.[72] Even in such a case, the randomization test with the order of entry of patients left unchanged would still be appropriate.[32,73]

The comparison of *P*-values between asymptotic and randomization tests in the simulated trials showed that on average, the absolute and relative discrepancy between tests is small and either test could be more powerful in a given dataset. For the smaller trials, the range of potential differences was wider, as expected. On average, the randomization *P*-values were found to be slightly lower for the larger simulated studies with 400 patients (right-hand side of Figure 2) and slightly higher for the smaller simulated studies with 100 patients (left-hand side of Figure 2). The randomization tests tended to be slightly more and slightly less conservative for the larger and smaller simulated trials, respectively. In the AMD trial, the minimization factors had little prognostic value and the above observation was unlikely to be attributed to the fact that stratified tests were used for trials of size 400, while unstratified tests were used for trials of size 100. It is more likely that the use of asymptotic tests was not warranted for the analysis of trials with only 25 patients per treatment arm. For treatment comparisons with *P*-values <0.05, the *P*-values of both tests were much more similar, with IQR of absolute differences equal to [−0.002, 0.003]. Thus, in situations where slight discrepancies could make the difference between a significant and a nonsignificant outcome, asymptotic and randomization tests largely concurred.

FDA prefers a randomization test for trials using minimization, as stated in the recent guidance on adaptive designs (2019): "*covariate-adaptive treatment assignment techniques do not directly increase the Type I error probability when analyzed with the appropriate methodologies (generally randomization or permutation tests)*".[47] Several authors have shown that minimization followed by a randomization test preserves the type-I error probability (see, e.g., Callegaro et al[63]).
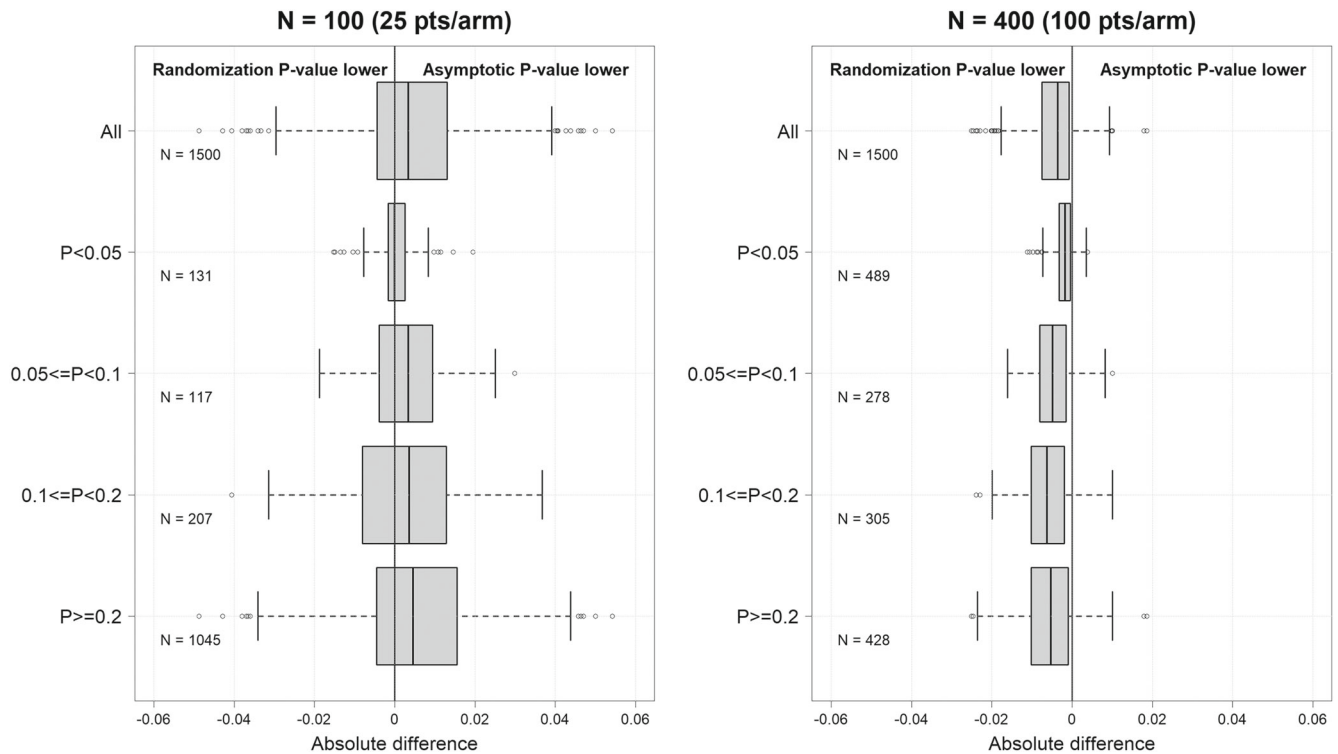
FIGURE 2 Absolute differences between asymptotic and randomization *P*-values based on 500 simulated trials, consisting of 4 arms and 3 comparisons each (total of 1500 *P*-values). Box plots of these absolute differences are shown for groupings by size of the asymptotic *P*-value.

In addition, minimization followed by an asymptotic test was also shown to preserve the probability and potentially be conservative depending on the stratification of the test.[11-14,63,64,74] Our results showed that in an actual clinical trial randomization and asymptotic tests provide similar inferences.

5 | IMPLEMENTATION AND REPORTING ISSUES

5.1 | Choice of parameters

Section 2.2.3. shows a simple minimization algorithm that is fit for purpose for most multicenter randomized trials. In this algorithm, the following choices must be made, based on clinical and statistical arguments, supported when necessary by simulations[75]:

- Prognostic factors: Only prognostic factors with a major impact on the outcomes of interest should be included. The choice of categories is either dictated by custom or chosen to make clinical sense. Although the number of prognostic factors and of categories do not matter much for minimization, it is pragmatically sensible to keep these numbers as small as possible.
- Center is not expected, in general, to be a modifier of the treatment effect, so the pursuit of a reasonable balance in each site stems mainly from a desire to avoid extreme imbalances within centers, if only because centers may occasionally need to be excluded from the analysis due to major bias, issues of trial conduct, and/or data quality.[76]
- When stratified blocked randomization is used, region is often included as one of the stratification factors. When minimization is used, it is preferable to include center as a minimization factor, in so far as region subsumes center. This will also ease the evaluation of data by regulatory and HTA agencies that will look at the treatment effects by country or region. As discussed in Section 3.5, not stratifying the analysis by center is not an issue, given the different objectives of using center during treatment allocation (for balance) and in the analysis (for power).

- "Study" as a factor. The inclusion of "study" in the minimization algorithm (counting the overall number of patients in each treatment group) may help achieve the overall balance (ie, the planned allocation ratio). For small studies, close attainment of the target sample sizes per group may be important.[27] Therefore, a randomization procedure in such studies should allow to closely match the desired allocation ratio and study could be included as factor.

- Weights for minimization factors: It is possible to assign different weights to the various prognostic factors in the minimization, in order to achieve the best balance for those factors thought to be of major relevance. Other procedures define a hierarchical list, and the imbalance at a lower level is not considered unless the imbalances at higher levels are under control.[50,51,77] Extensions of the minimization procedure exist which consider a weighted average of three types of imbalances: within stratum, within-covariate-margin, and overall. Note that the above-mentioned inclusion of the overall balance (with "study" as minimization factor) in the algorithm is related to this idea. In our experience, weighting is not needed to obtain good marginal balance for all factors. None of the 50 trials evaluated in Section 3.6 used weighting in the minimization algorithm.

- Bias used for biased-coin allocations; A value of β in the range of 0.3 to 0.4 seems sensible and has been used in registrational trials, while deterministic minimization is not recommended and would often not be accepted by regulators. ICH E9 recommends that if minimization is to be performed, b should be chosen <0.5 (or assignment to preferred treatment with probability <1).[42] Also, FDA guidance on adaptive trials mentions that "*predictability can be mitigated with an additional random component to prevent perfectly deterministic treatment assignment*".[47] All 50 trials discussed in Section 3.6 used a stochastic minimization algorithm.

- Function used to minimize imbalance: The variance method is commonly used, as it is simple to implement (Section 2.2.3.). For other choices of functions, see Pocock and Simon.[45]

- Missing values: In most situations, no missing values will be allowed for the minimization factors at the time of randomizing a patient. If a missing value is likely to occur for some of the factors, then it is advisable to add a factor level to allow for this possibility. As an example, in a trial for patients with community-acquired pneumonia, the pathogenic origin of pneumonia being bacterial or viral has a large impact on the outcomes of interest. However, for some patients, the pathogen may not yet be identified at the time of randomization, hence a category "unknown" may prove to be useful in order not to lose these patients. These are rare situations, and it is generally advisable to use only minimization factors that are completely known at the time of randomization.

5.2 | Unequal allocation ratios

Unequal allocation ratios can cause serious problems when minimization is used.[10] A "naïve" minimization algorithm can lead to allocation ratios that differ substantially from the target ratio. A modified minimization algorithm known as "biased-coin minimization" was proposed to achieve any allocation ratio.[52] However, this algorithm was criticized because it does not use the desired allocation ratio at every allocation step,[78] which can cause the unconditional randomization distribution of a test statistics to be shifted away from 0, causing low power of the randomization test and problems when interpreting study results.[10,78] An algorithm that keeps the allocation ratio at every allocation step, known as "allocation-ratio-preserving, biased-coin minimization", is essentially equivalent to creating as many virtual treatment groups as needed to obtain the desired allocation ratio, and then combining identical treatment groups.[78] For instance, if patients need to be randomized in a 2:1 allocation ratio between two treatment groups A and B, one can simply randomize patients equally between three virtual groups A1, A2, and B (ie, using a 1:1:1 allocation ratio), and then combine virtual treatment groups A1 and A2 to form a single actual treatment group A. If the algorithm creates treatment balance in each of three virtual groups, balance will also be obtained in the two actual treatment groups.

5.3 | Confidence intervals when using randomization tests

In this section, we briefly touch upon the randomization-based confidence interval estimation. More details can be found in the recent tutorial by Wang and Rosenberger.[79]

In contrast to a classical population-based approach, the one-to-one correspondence between hypothesis tests and confidence intervals does not exist for a randomization approach. Instead, the confidence interval can be computed using a whole set of related tests, and the mathematical model of the treatment difference is introduced without involving

an assumed population distribution.[79] Moving from hypothesis testing to confidence interval estimation, the *P*-value is evaluated across many sizes of treatment effect, say Δ, a treatment difference that is postulated for all patients in the study. The relationship between *P*-values and the coverage probability of the confidence interval is then utilized. For a constant additive effect Δ, the confidence interval from randomization tests is a set of Δ values for which the hypothesis HΔ is not rejected at the prescribed significance level.[80,81] Under HΔ, the probability of rejecting HΔ is understood as a generalized version of the type-I error rate.[82] Thus, the set of Δ values for which HΔ would not be rejected at level α can be regarded as a $100(1 - \alpha)\%$ interval estimate for the treatment difference. Wang and Rosenberger propose two approaches to estimate the boundaries of the randomization interval[79]: a stochastic approximation method (the Robbins-Monro algorithm, adapted from Garthwaite[83]) and a numerical approximation method (the bisection method).

5.4 | Validation and documentation

Minimization is often offered as part of a computerized (web-based) randomization and trial supply management system. Software developers typically configure the system using standard validated modules to adapt it to the specific requirements of a trial. A validation specialist validates the system using approved test scripts, and a biostatistician validates the minimization algorithm according to the trial protocol.[84] Once the system is operational, it is good practice to have an unblinded biostatistician performing monthly checks to ensure the randomization balance is correct. This holds true for any randomization algorithm.

Regulatory agencies have issued general guidance documents regarding computer system validation.[85,86] The European Medicines Agency also has a specific reflection paper on the use of interactive response technologies.[87] The guidance includes considerations about:

- accessibility of the system (ideally 24 hours a day, 7 days a week)
- definition of access permissions
- procedures for emergency unblinding (usually in the exceptional circumstance for which the patient safety requires the treatment to be known to the investigator)
- disaster recovery system, and description of back-up systems in case of power or Internet failures

  - manual interventions if required, along with appropriate documentation[88]
  - it is particularly important to put in place clear procedures for handling randomization errors, as these occur even in the best controlled environments. Useful guidance in this respect is provided by Yelland et al.[89]

- provision of a readily accessible, protected audit trail

It is good practice for the statistician programming the randomization test to first reproduce the actual treatment allocation in the statistical analysis software, with the same random seed used in the minimization software. This ensures that the programmed randomization procedure matches the actual allocation procedure.

5.5 | Recommendations for reporting

The method of treatment allocation is of vital importance in randomized trials, yet this aspect of the trial design is usually mentioned in a cursory manner, if at all, in published reports.[90] The validity of the procedure used to allocate treatments to patients can only be assessed indirectly from a trial's reported results. A comparison of treatment groups with respect to important prognostic factors, for instance, can reveal the presence of accidental or selection bias. For instance, treatment allocation in the SYMPHONY trial was criticized by a vigilant reader who observed that "*the imbalances observed are incompatible with the method described*".[91] In their reply, the authors explained that the imbalance observed was due to the fact that patients had been randomized in multiple centers, arguably an important design feature that was not mentioned in their published report.[92] Presence of imbalances does not automatically imply bias, nor does absence of observed imbalances rules out bias. This important topic is covered in great detail in the book by Berger.[31]

When balanced or dynamic randomization procedures are used, details must be given on the prognostic factors considered in the randomization, the algorithm used, and all other implementation details so that the reader can reconstruct the process as fully as possible, following CONSORT recommendations.[46] Insufficient details are likely to induce criticism or skepticism. For minimization, the essential details to be disclosed are the factors included in the minimization, the cut-off values used to categorize continuous variables, the criterion used to minimize overall imbalance (range or variance), the bias β used to favor the treatment group(s) minimizing the imbalance criterion (see Section 2.2.3 for details), and further details used to address specific situations, for example, unequal allocation ratios. This said, in order to avoid predictability of future treatment allocations, most details on the exact algorithm should not be documented in the trial protocol but in a separate document kept concealed from the study team and investigators. For the same reason, all information on study progress by treatment arm (such as numbers of patients by treatment arm in each center, overall or within stratification factor levels) should not be shared with study the team and investigators. Without sharing of such information, it is highly unlikely that any investigator in a multicenter trial will have access to all information needed to predict the allocation probabilities for the next patient.

5.6 | Open software for minimization

OxMAR is an online system that can be accessed from anywhere using any device which can look at a web page.[93] The system uses a series of Perl scripts. The result of the allocation procedure (minimization if desired) is stored in a text file, is output to a webpage and is emailed to the user and study administrator. Further details are available at https://sourceforge.net/projects/oxmar/. MinimPy is a free, open-source desktop application for "managing the whole process of minimization". The MinimPy program is written using Python. The program is freely available for download under GPL license. Further details are available at https://pypi.org/project/MinimPy/. The R-package Minirand generates randomization schedules for two or more treatment arms and any allocation ratio by minimization algorithms.[94] The package can be freely downloaded at https://cran.r-project.org/web/packages/Minirand/index.html. The recent R-package carat supports six covariate-adaptive randomization procedures and three hypothesis-testing methods. Additionally, the package provides efficient tools for evaluation and comparison of the performance of randomization procedures and tests under different assumptions. The package can be freely downloaded at https://cran.r-project.org/web/packages/carat/index.html.

6 | DISCUSSION

Table 9 summarizes the advantages and disadvantages of minimization, which indeed does an excellent job at achieving balance across many factors simultaneously. The main reason for aiming at such balance is to increase the perceived scientific validity of the study outcome by stakeholders, including regulatory and HTA agencies. All procedures aimed at improving balance make a tradeoff, and the nature of the tradeoff depends on the study characteristics (use of masking, knowledge of baseline covariates, etc.). Although balanced treatment allocation is not strictly necessary (see Section 2.1.4), a study with accidental confounding (even if with correctly implemented randomization) will be unsatisfactory in many respects, as pointed out by Efron.[53] This has indeed been confirmed in actual clinical trials.[95] As for all restricted randomization schemes, the predictability of the next allocation will be larger with minimization than with completely random allocation if details about enrolled patients are shared and caution should be taken to avoid potential selection bias. However, predictability and potential selection bias should not be an issue in well-conducted multicenter trials. For these reasons, the trade-off between treatment balance and predictability is most often in favor of achieving balance in well controlled multicenter trials.[96]

Flexibility is a clear advantage of minimization that is important in practice: the procedure "self-adapts" to any situation in terms of number of prognostic factors, number of categories for these factors, target sample size, and center-specific sample sizes, which are usually hard to predict. This may explain why minimization is in common use in oncology, where randomized trials considerably vary in size (from a few dozen patients in phase 2 trials to a few thousand in phase 3), and in the number and size of participating centers. In contrast, stratified blocked allocation requires careful assessment of the number of factors that can be stratified for, to avoid partially filled blocks and overall imbalances. Minimization can be considered a valid alternative to permuted-block randomization when the number of strata becomes too large. For small trials, minimization can be an advantageous alternative to simple randomization. Importantly, stochastic minimization

TABLE 9 Main advantages and disadvantages of minimization.

Advantages	Disadvantages
• Ensures good balance for possibly many prognostic factors	• Balances over prognostic factor margins and not over strata.
• Does not generate any deterministic treatment allocation (unlike permuted block designs)	• Dynamic method, hence randomization lists cannot be generated in advance and requires software for implementation
• Self-adapts to any situation in terms of number of factors and/or sample sizes	• Requires a randomization test for the analysis
• Can use continuous covariates for balancing	

does not have deterministic treatment allocations, whereas permuted blocks within strata lead to some deterministic allocations. This flexibility includes the use of continuous covariates for balancing in addition to factors. Recently, the theoretical framework for inference when directly handling continuous variables for balancing has been explored, which may lead to implementing minimization procedures with both categorical and continuous variables.[14,97]

The flexibility of the minimization algorithm comes at the cost of requiring a software to implement the dynamic treatment allocation. For unequal allocation ratios, it is important to apply a procedure that uses the desired allocation ratio at every allocation step, and "allocation-ratio-preserving, biased-coin minimization" is a simple procedure satisfying this condition.

The theoretical framework for inference after minimization was established, demonstrating that asymptotic tests control the type-I error probability, and can even be conservative if the analysis is not stratified for all balancing factors.[11-13] In practice, randomization tests may still be required to assuage any concerns about the validity of the treatment-allocation procedure when minimization is used. Nowadays, performing several thousands of simulations (rerandomizations) for a randomization test is easily computationally feasible. As shown here and in other studies, such simulations have shown both preservation of the type-I error probability and good power of randomization tests, for trials of different sizes and with different number of minimization factors and levels.[63,98] In addition, a randomization test based on the random assignment model[99] preserves the type-I error probability in the presence of time trends,[73,100] whereas asymptotic tests do not control the probability in case of time-drift for which no adjustment is made in the analysis. Moreover, randomization-based inference is valid in a minimization context independently of a correctly specified model.[1] All these properties render the randomization test a sound primary analysis method for clinical trials using minimization. As argued by Berger, randomization-based tests are in general robust and valid alternatives to likelihood-based tests and should be considered more frequently by clinical investigators.[27] Of note, the choice of a test should be based on the same reasoning whether minimization is used or any other type of restricted design (eg, stratified permuted block design). In practice, asymptotic tests are used uncritically regardless of the treatment allocation method. We and others have shown that asymptotic tests safeguard the type-I error probability for trials using minimization[11-13,63,64,75] so the practice is not expected to increase the number false-positive trials.

A concern sometimes expressed with minimization is that by forcing balance on a few factors used in the procedure, accidental bias can be created for factors not used (including unknown factors which by definition cannot be used). Taves argued that this concern is unfounded.[44] Aickin elaborated on Taves' argument that minimization could only produce *better* balance, on average, than completely random allocation: it would have no effect on factors uncorrelated to the factors used in minimization, and it would tend to produce better balance for factors correlated to the factors used.[66] This was theoretically shown recently[67] and hence the balance of factors not used in the allocation procedure is no longer of concern for the use of minimization.

Minimization is versatile and can be used advantageously in many situations, but it is especially useful when (1) many baseline factors are known to have a major impact on the patient's prognosis, (2) many centers accrue patients into the trial, and the centers may contribute very different numbers of patients, or (3) the trial sample size is small (less than 100 patients) to medium (a few 100 patients). If any of these three conditions holds, minimization may be preferred to simple randomization, or to stratified blocked allocation. As Aickin stated: "*In small studies, or large studies with many important prognostic covariates or analytic subgroups, design-adaptive allocation is an extremely attractive method of treatment assignment*".[66]

ORCID

Elisabeth Coart https://orcid.org/0000-0002-0018-1919

REFERENCES

1. Rosenberger WF, Lachin JM. *Randomization in Clinical Trials: Theory and Practice.* 2nd ed. Hoboken: John Wiley & Sons; 2016.
2. Senn S. Testing for baseline balance in clinical trials. *Stat Med*. 1994;13(17):1715-1726.
3. Hamilton SA. Dynamically allocating treatment when the cost of goods is high and drug supply is limited. *Control Clin Trials*. 2000;21(1):44-53.
4. McEntegart D. Forced Randomization: When Using Interactive Voice Response Systems. Accessed 13 March 2022, https://www.appliedclinicaltrialsonline.com/view/forced-randomization-when-using-interactive-voice-response-systems
5. Altman DG, Bland JM. Treatment allocation by minimisation. *BMJ*. 2005;330(7495):843.
6. European Medicines Agency. Points to consider on adjustment for baseline covariates. Accessed 21 March 2022: https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-adjustment-baseline-covariates_en.pdf
7. Senn S, Anisimov VV, Fedorov VV. Comparisons of minimization and Atkinson's algorithm. *Stat Med*. 2010;29(7–8):721-730.
8. Senn S. Seven myths of randomisation in clinical trials. *Stat Med*. 2013;32(9):1439-1450.
9. Berger VW. Minimization: not all it's cracked up to be. *Clin Trials*. 2011;8(4):443.
10. Proschan M, Brittain E, Kammerman L. Minimize the use of minimization with unequal allocation. *Biometrics*. 2011;67(3):1135-1141.
11. Shao J, Yu X. A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*. 2010;97(2):347-360.
12. Shao J. Inference after covariate-adaptive randomisation: aspects of methodology and theory. *Stat Theory Relat Fields*. 2021;5(3):172-186.
13. Ma W, Hu F, Zhang L. Testing hypotheses of covariate-adaptive randomized clinical trials. *J Am Stat Assoc*. 2015;110(510):669-680.
14. Ma W, Qin Y, Li W, Hu F. Statistical inference for covariate-adaptive randomization procedures. *J Am Stat Assoc*. 2020;115(531):1488-1497.
15. Kaiser LD. Dynamic randomization and a randomization model for clinical trials data. *Stat Med*. 2012;31(29):3858-3873.
16. Ohashi Y. Randomization in cancer clinical trials: permutation test and development of a computer program. *Environ Health Perspect*. 1990;87:13-17.
17. McPherson GC, Campbell MK, Elbourne DR. Use of randomisation in clinical trials: a survey of UK practice. *Trials*. 2012;13:198.
18. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N Engl J Med*. 2005;353(16):1659-1672.
19. Green H, McEntegart DJ, Byrom B, Ghani S, Shepherd S. Minimization in crossover trials with non-prognostic strata: theory and practical application. *J Clin Pharm Ther*. 2001;26(2):121-128.
20. ISIS-3: a randomised comparison of streptokinase vs tissue plasminogen activator vs anistreplase and of aspirin plus heparin vs aspirin alone among 41,299 cases of suspected acute myocardial infarction. ISIS-3 (third international study of infarct survival) collaborative Group. *Lancet*. 1992;339(8796):753-770.
21. Lachin JM, Matts JP, Wei LJ. Randomization in clinical trials: conclusions and recommendations. *Control Clin Trials*. 1988;9(4):365-374.
22. Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. *Control Clin Trials*. 1988;9(4):327-344.
23. McEntegart DJ. Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial by Christopher J. Weir and Kennedy R. Lees, statistics in medicine 2003; 22:705-726. *Stat Med*. 2004;23(23):3719-3720. author reply 3720-3711.
24. Hedden SL, Woolson RF, Malcolm RJ. Randomization in substance abuse clinical trials. *Subst Abuse Treat Prev Policy*. 2006;1:6.
25. 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.
26. Hu F, Hu Y, Ma Z, Rosenberger WF. Adaptive randomization for balancing over covariates. *WIREs Comput Stat*. 2014;6(4):288-303.
27. Berger VW, Bour LJ, Carter K, et al. A roadmap to using randomization in clinical trials. *BMC Med Res Methodol*. 2021;21(1):168.
28. Stigsby B, Taves DR. Rank-minimization for balanced assignment of subjects in clinical trials. *Contemp Clin Trials*. 2010;31(2):147-150.
29. Taves DR. Rank-minimization with a two-step analysis should replace randomization in clinical trials. *J Clin Epidemiol*. 2012;65(1):3-6.
30. Lebowitsch J, Ge Y, Young B, Hu F. Generalized multidimensional dynamic allocation method. *Stat Med*. 2012;31(28):3537-3544.
31. Berger VW. *Selection Bias and Covariate Imbalances in Randomized Clinical Trials.* Chichester: Wiley & Sons; 2005.
32. Frane JW. A method of biased coin randomization, its implementation, and its validation. *Drug Inf J*. 1998;32:423-432.
33. Robertson DS, Lee KM, Lopez-Kolkovska BC, Villar SS. Response-adaptive randomization in clinical trials: from myths to practical considerations. arXiv:2005.00564 [stat.ME].
34. Begg CB, Iglewicz B. A treatment allocation procedure for sequential clinical trials. *Biometrics*. 1980;36(1):81-90.
35. Atkinson AC. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*. 1982;69(1):61-67.
36. Begg CB, Kalish LA. Treatment allocation for nonlinear models in clinical trials: the logistic model. *Biometrics*. 1984;40(2):409-420.
37. Atkinson AC. Optimum biased-coin designs for sequential treatment allocation with covariate information. *Stat Med*. 1999;18(14):1741-1752. discussion 1753–1745.

38. Mavrogonatoua L, Sun Y, Robertson DS, Villar SS. Comparison of allocation strategies for optimising clinical trial designs under variance heterogeneity. *Comput Stat Data Anal*. 2022;176:176. doi:10.1016/j.csda.2022.107559

39. Zhao W, Berger V. Imbalance control in clinical trial subject randomization-from philosophy to strategy. *J Clin Epidemiol*. 2018;101:116-118.

40. Rosenberger WF, Sverdlov O. Handling covariates in the design of clinical trials. *Stat Sci*. 2008;23(3):404-419.

41. Berger VW, Ivanova A, Knoll MD. Minimizing predictability while retaining balance through the use of less restrictive randomization procedures. *Stat Med*. 2003;22(19):3017-3028.

42. European Medicines Agency. ICH Topic E9. Statistical Principles for Clinical Trials. 1998 Accessed 13 March 2022: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf

43. Therneau TM. How many stratification factors are "too many" to use in a randomization plan? *Control Clin Trials*. 1993;14(2):98-108.

44. Taves DR. Minimization: a new method of assigning patients to treatment and control groups. *Clin Pharmacol Ther*. 1974;15(5):443-453.

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

46. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Trials*. 2010;11:32.

47. U.S. Department of Health and Human Services. Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry. 2019 Accessed 14 March 2022: https://www.fda.gov/media/78495/download

48. Freedman LS, White SJ. On the use of Pocock and Simon's method for balancing treatment numbers over prognostic factors in the controlled clinical trial. *Biometrics*. 1976;32(3):691-694.

49. White SJ, Freedman LS. Allocation of patients to treatment groups in a controlled clinical study. *Br J Cancer*. 1978;37(5):849-857.

50. Signorini DF, Leung O, Simes RJ, Beller E, Gebski VJ, Callaghan T. Dynamic balanced randomization for clinical trials. *Stat Med*. 1993;12(24):2343-2350.

51. Heritier S, Gebski V, Pillai A. Dynamic balancing randomization in controlled clinical trials. *Stat Med*. 2005;24(24):3729-3741.

52. Han B, Enas NH, McEntegart D. Randomization by minimization for unbalanced treatment allocation. *Stat Med*. 2009;28(27):3329-3346.

53. Efron B. International Symposium on Hodgkin's Disease. Session 6. Survival data and prognosis. Invited discussion: forcing a sequential experiment to be balanced. *Natl Cancer Inst Monogr*. 1973;36:571-572.

54. Cyclophosphamide plus cisplatin versus cyclophosphamide, doxorubicin, and cisplatin chemotherapy of ovarian carcinoma: a meta-analysis. The ovarian cancer meta-analysis project. *J Clin Oncol*. 1991;9(9):1668-1674.

55. Blackwell D, Hodges JJ Jr. Design for the control of selection bias. *Ann Math Stat*. 1957;28(2):449-460.

56. Hagino A, Hamada C, Yoshimura I, Ohashi Y, Sakamoto J, Nakazato H. Statistical comparison of random allocation methods in cancer clinical trials. *Control Clin Trials*. 2004;25(6):572-584.

57. Toorawa R, Adena M, Donovan M, Jones S, Conlon J. Use of simulation to compare the performance of minimization with stratified blocked randomization. *Pharm Stat*. 2009;8(4):264-278.

58. Xiao L, Lavori PW, Wilson SR, Ma J. Comparison of dynamic block randomization and minimization in randomized trials: a simulation study. *Clin Trials*. 2011;8(1):59-69.

59. Hills RK, Gray R, Wheatley K. Balancing treatment allocations by clinician or center in randomized trials allows unacceptable levels of treatment prediction. *J Evid Based Med*. 2009;2(3):196-204.

60. Brown S, Thorpe H, Hawkins K, Brown J. Minimization—reducing predictability for multi-Centre trials whilst retaining balance within Centre. *Stat Med*. 2005;24(24):3715-3727.

61. Perry M, Faes M, Reelick MF, Olde Rikkert MG, Borm GF. Studywise minimization: a treatment allocation method that improves balance among treatment groups and makes allocation unpredictable. *J Clin Epidemiol*. 2010;63(10):1118-1122.

62. Chen LH, Lee WC. Two-way minimization: a novel treatment allocation method for small trials. *PLoS One*. 2011;6(12):e28604.

63. Callegaro A, Harsha Shree BS, Karkada N. Inference under covariate-adaptive randomization: a simulation study. *Stat Methods Med Res*. 2021;30(4):1072-1080.

64. Birkett NJ. Adaptive allocation in randomized controlled trials. *Control Clin Trials*. 1985;6(2):146-155.

65. Liu H, Tu F, Ma W. Lasso-adjusted treatment effect estimation under covariate-adaptive randomization. *Biometrika*. 2023;110(2):431-447.

66. Aickin M. Randomization, balance, and the validity and efficiency of design-adaptive allocation methods. *J Stat Plan Inference*. 2001;94:97-109.

67. Liu Y, Hu F. Balancing unobserved covariates with covariate-adaptive randomized experiments. *J Am Stat Assoc*. 2022;117(538):875-886.

68. Wang Y, Rosenberger WF, Uschner D. Randomization-based inference and the choice of randomization procedures. *Stat Pap*. 2019;60:395-404.

69. Onghena P. Randomization tests or permutation tests? A historical and terminological clarification. In: Berger V, ed. *Randomization, Masking, and Allocation Concealment*. Boca Raton: Chapman & Hall/CRC Press; 2018:209-227.

70. Gragoudas ES, Adamis AP, Cunningham ET Jr, Feinsod M, Guyer DR, Group VISiONCT. Pegaptanib for neovascular age-related macular degeneration. *N Engl J Med*. 2004;351(27):2805-2816.

71. Edgington E, Onghena P. *Randomization Tests*. 4th ed. New York: Chapman and Hall/CRC; 2007.

72. Kalish LA, Begg CB. The impact of treatment allocation procedures on nominal significance levels and bias. *Control Clin Trials*. 1987;8(2):121-135.

73. Simon R, Simon NR. Using randomization tests to preserve type I error with response-adaptive and covariate-adaptive randomization. *Stat Probab Lett*. 2011;81(7):767-772.

74. Barbáchanoa Y, Coad DS, Robinson DR. Predictability of designs which adjust for imbalances in prognostic factors. *J Stat Plan Inference*. 2008;138:756-767.

75. Wade A, Pan H, Eaton S, Pierro A, Ong E. An investigation of minimisation criteria. *BMC Med Res Methodol*. 2006;6:11.

76. Herson J. Strategies for dealing with fraud in clinical trials. *Int J Clin Oncol*. 2016;21(1):22-27.

77. Hu Y, Hu F. Asymptotic properties of covariate-adaptive randomization. *Ann Stat*. 2012;40(3):1794-1815.

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

79. Wang Y, Rosenberger WF. Randomization-based interval estimation in randomized clinical trials. *Stat Med*. 2020;39(21):2843-2854.

80. Michiels B, Heyvaert M, Meulders A, Onghena P. Confidence intervals for single-case effect size measures based on randomization test inversion. *Behav Res Methods*. 2017;49(1):363-381.

81. Onghena P. Randomization and the randomization test: two sides of the same coin. In: Berger V, ed. *Randomization, Masking, and Allocation Concealment*. Boca Raton: Chapman & Hall/CRC Press; 2018:185-207.

82. Kempthorne O. Why randomize? *J Stat Plan Inference*. 1977;1:1-25.

83. Garthwaite PH. Confidence intervals from randomization tests. *Biometrics*. 1996;52:1387-1393.

84. Downs M, Tucker K, Christ-Schmidt H, Wittes J. Some practical problems in implementing randomization. *Clin Trials*. 2010;7(3):235-245.

85. U.S. Department of Health and Human Services. Food and Drug Administration. General Principles of Software Validation; Final Guidance for Industry and FDA Staff. 2002 Accessed 16 March 2022: https://www.fda.gov/media/73141/download

86. European Medicines Agency. Guideline on computerised systems and electronic data in clinical trials. 2021 Accessed 16 March 2022: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/draft-guideline-computerised-systems-electronic-data-clinical-trials_en.pdf

87. European Medicines Agency. Reflection paper on the use of interactive response technologies (interactive voice/web response systems) in clinical trials, with particular emphasis on the handling of expiry dates. 2013 Accessed 16 March 2022: https://www.ema.europa.eu/en/documents/scientific-guideline/superseded-reflection-paper-use-interactive-response-technologies-interactive-voice/web-response-systems-clinical-trials-particular-emphasis-handling-expiry-dates_en.pdf

88. Cai HW, Xia JL, Gao DH, Cao XM. Implementation and experience of a web-based allocation system with Pocock and Simon's minimization methods. *Contemp Clin Trials*. 2010;31(6):510-513.

89. Yelland LN, Sullivan TR, Voysey M, Lee KJ, Cook JA, Forbes AB. Applying the intention-to-treat principle in practice: guidance on handling randomisation errors. *Clin Trials*. 2015;12(4):418-423.

90. Williams DH, Davis CE. Reporting of assignment methods in clinical trials. *Control Clin Trials*. 1994;15(4):294-298.

91. Altman DG. Randomisation in the SYMPHONY trial. *Lancet*. 2000;356(9240):1521-1522.

92. Comparison of sibrafiban with aspirin for prevention of cardiovascular events after acute coronary syndromes: a randomised trial. The SYMPHONY InvestigatorsSibrafiban versus aspirin to yield maximum protection from ischemic heart events post-acute coronary syndromes. *Lancet*. 2000;355(9201):337-345.

93. O'Callaghan CA. OxMaR: open source free software for online minimization and randomization for clinical trials. *PLoS One*. 2014;9(10):e110761.

94. Jin M, Polis A, Hartzel J. Algorithms for minimization randomization and the implementation with an R package. *Commun Stat Simul Comput*. 2019;50:3077-3087. doi:10.1080/03610918.2019.1619765

95. Verberk WJ, Kroon AA, Kessels AG, et al. Comparison of randomization techniques for clinical trials with data from the HOMERUS-trial. *Blood Press*. 2005;14(5):306-314.

96. McPherson GC, Campbell MK, Elbourne DR. Investigating the relationship between predictability and imbalance in minimisation: A simulation study. *Trials*. 2013;14:86.

97. Ma W, Li P, Zhang L-X, Hu F. A new and unified family of covariate adaptive randomization procedures and their properties. *J Am Stat Assoc*. 2022;1-12. doi:10.1080/01621459.2022.2102986

98. Hasegawa T, Tango T. Permutation test following covariate-adaptive randomization in randomized controlled trials. *J Biopharm Stat*. 2009;19(1):106-119.

99. Kempthorne O, Doerfler TE. The behavior of some significance tests under experimental randomization. *Biometrika*. 1969;56:231-248.

100. Wei LJ, Lachin JM. Properties of the urn randomization in clinical trials. *Control Clin Trials*. 1988;9(4):345-364.