École nationale de la statistique
et de l'analyse de l'information

Internship report

Statiscal properties of Minimal Sufficient Balance, Minimization and Stratified Permuted Blocks in a stratification context, a simulation study

Hugo Cannafarina

Tutors:

ARNOUX Armelle

MURRIS Juliette

2023

Remerciements

Je tiens à remercier mes tutrices, Armelle Arnoux et Juliette Murris, pour tout le temps et la disponibilité qu'elles m'ont accordés tout au long de ce projet. Le sujet proposé était extrêmement intéressant, et réfléchir avec elles à comment surmonter tous les problèmes recontrés était extrèmement agréable.

Je tiens aussi à remercier Maud Megret, qui avait un sujet très proche du mien, pour tous les moments d'entraides et les conversations stimulantes.

Merci également à Emilien, Julien, Julie, Hélène et Ilona pour la très bonne ambiance qui régnait dans l'équipe des stagiaires, et sans qui les journées auraient été plus mornes.

Merci à François Hardy pour le temps dédié aux stagiaires et à tous les enseignments très instructifs sur les essais cliniques.

Merci aussi à Sébastien Da-Veiga pour avoir lu ma note d'étape intermédiaire et pour le temps qu'il va accorder à la correction de ce rapport.

Merci à Tissine pour le soutien et l'aide à la relecture.

Enfin, je remercie tout le reste des équipes de l'Unité de Recherche Clinique de Georges Pompidou et de l'équipe HEKA pour l'accueil chaleureux qui m'a été fait, et toutes les personnes qui m'ont soutenu lors de ce projet, de quelque manière que ce soit.

Environnement d'accueil et objectifs

J'ai effectué mon stage dans les locaux de Paris Santé Campus, au sein de l'équipe HEKA qui a pour vocation de développer des méthodologies et des outils à partir de concepts mathématiques, statistiques et informatiques avancés afin d'aider à la décision clinique. Cependant, j'étais rattaché au module « Epidémiologie Clinique » (EC) de l'Unité de Recherche Clinique (URC) de l'Hôpital George Pompidou. Le module EC a pour mission de mener des recherches cliniques dans différents domaines thématiques. Il collabore avec des cliniciens et des chercheurs pour les aider à formuler leurs questions de recherche, élaborer les protocoles d'études, coordonner les projets, gérer les données et effectuer des analyses statistiques.

Ce stage avait pour but d'étudier divers algorithmes de randomisation adaptative afin d'aider les chercheurs de l'URC à les implémenter dans de vrais essais cliniques. Les chercheurs à l'origine de l'essai clinique Sepsiscool-2, qui fait usage d'une procédure de randomisation adaptative, ont besoin de mieux comprendre le fonctionnement et les avantages de cette méthode pour mener à bien des analyses statistiques plus rigoureuses. Cet article a aussi pour but de faciliter la rédaction d'un article sur la comparaison de procédures de randomisation adaptative qui sera soumis pour publication. C'est pourquoi il a été écrit en anglais, et une attention particulière a été portée au respect des codes de la recherche en statistiques. Le code utilisé pour les simulation effectuées dans ce rapport peut être retrouvé sur ma page github : `https://github.com/HugoCannafarina/AdaptativeRandomisationInternship`.

Abstract

Background : In clinical trials, it is essential to ensure that patients in both groups have similar baseline characteristics. To achieve this, covariate-adaptive randomization methods can be used to control for balance. Minimization is the most common of these randomization techniques, but has been criticized. When using this procedure, continuous covariates need to be discretized, and the proportion of non-random patient assignment is high. Minimal Sufficient Balance is a relatively new covariate-adaptive randomization procedure that has been developed to overcome these drawbacks. The aim of this article is to compare stratified permuted block randomization, the most common randomization technique, Minimization and Minimal Sufficient Balance on data from a clinical trial in intensive care, using as of yet untested criteria.

Methods : First, the Sepsicool-1 dataset was augmented using the NORTA method, which allows to generate correlated datasets with any specified distribution. Four different scenarios were considered to vary the sample sizes and correlation structures of the datasets. 5000 datasets were generated for each scenario. Three algorithms, Stratified Permuted Blocks, Minimization and Minimal Sufficient Balance, were then applied to each dataset with stratification on at least one covariate. Stratified and unstratified imbalance tests (Student's t-test and Wilcoxon signed-rank test) were performed, and the ability of each procedure to prove treatment efficacy was assessed.

Results : No randomization procedure was clearly better than the others at proving treatment's effectiveness. However, Minimal Sufficient Balance outperformed the other algorithms at minimizing imbalance with stratified Students-t-test and stratified Wilcoxon signed-rank test, as well as non-stratified Students-t-test. Minimization was better for non-stratified Wilcoxon signed-rank test

Conclusion : In most cases, Minimal Sufficient Balance is an excellent choice for showing that a treatment is effective while minimizing group imbalance. If after stratification, a non-stratified analysis is to be performed, Minimization may be slightly more effective in reducing imbalance, particularly if the patients' baseline covariates are not normally distributed. Even in this case, Minimal Sufficient Balance may also be a good choice, as it presents better levels of allocation randomness.

Summary

1 Introduction

The goal of a clinical trial is to prove the efficiency of a treatment by comparing it to a control treatment. The comparison can only make sense if the patients to which each treatment is administered share the same characteristics. If they don't, we speak of group imbalance or covariate imbalance, in which case the results risk being unusable.

Randomization is the process which consists in randomly allocating patients to one treatment or another. Randomization should prevent any prior knowledge of group assignment. It is crucial because in an unblinded clinical trial (where the researchers know which treatment is administered), the ability of researchers to predict upcoming treatment allocations can lead to biased patient selection based on individual characteristics. For instance, in a trial seeking to determine the most effective treatment for reducing mortality, if researchers possess a preference for one specific treatment, they might choose to administer it solely to severely ill patients. This introduces selection bias, which causes covariate imbalance.

In theory, since patients in each treatment group are randomly selected, randomization should also ensure balance between the groups. In practice however, group imbalance can be observed even if randomization is done properly. The most common statistical tests used to assess covariate imbalance are the chi-squared test for categorical variables and Student's t-test or Wilcoxon signed-rank test for quantitative variables. These tests establish whether or not the two groups of patients come from the same population, which should be the case if there is no selection bias. However, some of these tests will be rejected even if there is no selection bias, by definition of the Type-I error. Since it is generally set at 5%, 1 out of 20 imbalance tests are rejected if nothing is done to ensure balance during randomization.

The purpose of the randomization algorithms studied in this article is to ensure balance on certain important covariates, while minimizing the risks of selection bias. The most commonly used methods are stratified randomization and covariate adapatative randomization. The main difference between these two approaches is that the treatment allocation is defined during the design phase of the trial in stratification, while it is defined depending on every previous allocation in covariate adaptative randomization. Covariate adaptative randomization is used when many covariates have to be balanced.

Stratified permuted blocks, theorized by Efron and Bradley [9], is a randomization method that ensures group balance for a certain amount of categorical variables and that limits the risks of selection bias by maintaining random assignments. Moreover, it guarantees a similar size for the two treatment groups (which is not guaranteed if the randomization simply consists in flipping a coin to determine the treatment allocation) and thus helps maximise the statistical power during the final analysis. Minimization (Pocock, Stuart J., and Richard Simon, 1975 [19]) and Minimal Sufficient Balance (Zhao, Wenle, Michael D. Hill, and Yuko Palesch, 2015 [24]) are covariate adaptative randomization methods that allow to balance over a large amount of covariates. They are called adaptative because the treatment allocation of each patient depends on the treatments and characteristics of every previous patients. Minimization has been criticized because it requires to transform continuous covariates into categorical ones and because of the very low level of allocation randomness.

Using data from a stroke study, S. D. Lauzon et al. [16] have shown that the p-values of imbalance tests are on average higher when using Minimal Sufficient Balance rather than Minimization. Moreover, the percentage of pure random assignments is higher in the first case, which shows the superiority of Minimimal Sufficient Balance at preserving allocation randomness. In addition, S. D. Lauzon et al. [15] have shown, using simulated data, that power levels in the estimation of treatment effect were similar when using complete randomization, Minimization, and Minimal Sufficient Balance.

The aim of this article is to compare the performance of the different algorithms in order to better understand their respective strengths. This article looks at the algorithms' ability to balance covariates, in particular when the Wilcoxon signed-rank test is used to measure the imbalance of continuous covariates and when the algorithms are stratified. To our knowledge, these particular aspects have not yet been studied in the literature. Moreover, the ability of the algorithms to show the effectiveness of the treatment is being investigated. The data used was generated from a real clinical trial in intensive care called Sepsiscool-1 [21] using the NORTA (NORmal To Anything) method, which allows to simulate correlated datasets with any specified distribution. This method allows realistic clinical trial datasets of any size to be obtained.

Section 2 precisely describes the algorithms, the method used to generate data and the criteria using which they will be compared. The results are presented in Section 3 and are discussed in Section 4.

2 Methodology

2.1 Randomization methods

2.1.1 Stratified permuted blocks

Suppose that a clinical trial consists of two treatments, A and B. A block is a group of consecutive participants that are randomly assigned to different treatment groups in a balanced way. For example, let's assume, without loss of generality, that the block size is set at four. Every four participants, two of the following participants will receive treatment A and two will receive treatment B in random order. The order could be $AABB$, for example, meaning that the first two patients will receive treatment A and the next two will receive treatment B. Blocks are said to be permuted because the order in which treatments are given in a block can be any permutation of $AABB$ ($ABAB$, $BABA$, $BBAA$...).

The advantage of using permuted blocks is to ensure that the size of the two treatment groups is roughly similar, thus maximizing statistical power in the final analysis. This would not necessarily be the case if randomization simply involved flipping a coin to determine treatment allocations. Moreover, the permutation procedure makes it more difficult for researchers to guess which allocation will come next, thus limiting the risk of selection bias.

The point of stratification is to ensure that certain patient characteristics are balanced within each group. For example, one can stratify on clinical center to ensure that there are as many people from each clinical center in both groups. Stratification is only possible on categorical covariates. It is possible to stratify on more than one covariate, in which case each combination of covariate levels is balanced between the groups.

In practice, with permuted block stratification, a block is made up of participants with the same combination of covariate levels. An example of randomization with blocks of size four and stratification on two covariates is given in table 1.

Covariate 1	Category 1		Category 2	
Covariate 2	Category 1	Category 2	Category 1	Category 2
Block number	Treatment allocation			
1	B	B	B	A
1	A	B	A	A
1	B	A	A	B
1	A	A	B	B
2	A	A	A	B
2	B	A	B	B
2	B	B	B	A
2	A	B	A	A
⋮	⋮	⋮	⋮	⋮

Table 1: Exemple of block randomization with blocks of size 4 and two covariates with two categories each

2.1.2 Minimal Sufficient Balance

This randomization procedure is initialized according to the random allocation rule. This rule is described by Rosenberger, William F., and John M. Lachin. in *Randomization in clinical trials: theory and practice* [20]. Its aim is to ensure that the size of the two groups is approximately similar after 20 patients:

- The first patient is assigned to treatment arm A with probability $1/2$.
- The patient number l is assigned to group A with probability $\frac{20/2-n_a}{20-l+1}$, where n_a is the number of patients previously allocated to treatment A

The treatment allocation is then based on votes. Let's assume that m covariates have to be balanced during the randomization process. For each patient, a vote is collected for each covariate. The vote indicates which treatment group the patient should be placed in to reduce the imbalance between groups for that covariate. The rule to decide which treatment to give to a new patient is the following :

- if there are more votes for treatment A than for treatment B, treatment A is given with proba-

bility ξ.

- if there are more votes for treatment B than for treatment A, treatment B is given with probability $1-\xi$.
- if there are as many votes for both treatment, treatment A is given with probability $1/2$.

ξ is called the biased coin probability. Its purpose is to keep an element of random in the treatment allocation, in order to avoid selection bias. The rule for obtaining a vote depends on the type of the covariate :

Suppose that the covariate is quantitative : consider n_a, $\bar{x}_a$ and s_a respectively the total number of subjects previously randomized in the treatment group A, the current mean of the covariate in this group and its current variance. Also consider n_b, $\bar{x}_b$ and s_b the equivalent statistics in treatment group B. Let t be the test statistic for a Welch test, described by Welch, Bernard L. [23].

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

This statistic follows a student law with degrees of freedom ν :

$$\nu = \frac{\left(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}\right)^2}{\frac{s_a^4}{n_a^2(n_a-1)} + \frac{s_b^4}{n_b^2(n_b-1)}}$$

The Welch test is similar to Student's t-test, but is more appropriate when the two populations have unequal variances and sizes. However, the test still assumes that the means of the population samples are normally distributed. Let p be the p-value of this test. Moreover, let p^* be the p-value threshold below which we consider there to be covariate imbalance and t^* be the test statistic corresponding to the imbalance threshold p^*. t^* is calculated with the distribution function of a student law with ν degrees of freedom. If x is the current subject's covariate value, then the voting law for the covariate is :

- Vote for A if ($t < -t^*$ and $x > \bar{x}_b$) or if ($t > t^*$ and $x < \bar{x}_b$)
- Vote for B if ($t < -t^*$ and $x < \bar{x}_a$) or if ($t > t^*$ and $x > \bar{x}_a$)
- Neutral otherwise

Thus, if the covariate's mean is significantly different in both groups (p $< p^*$) and the current patient's covariate value is not between these means, the vote is cast in favor of the treatment group in which the patient would contribute to reducing the difference between the means.

Suppose that the covariate is categorical (with a small amount of categories) : Define g the number of categories of the covariate. For $1 \leq j \leq g$ define n_{ja} the number of subjects in the category j previously randomized to treatment A and n_{jb} the same statistic for treatment B. The theoretical number of patient in the j category for both treatment is equal to :

$$E_{jk} = \frac{\left(n_{ja}+n_{jb}\right)\sum_{i=1}^{g} n_{ik}}{\sum_{i=1}^{g}(n_{ia}+n_{ib})}, \quad (k = a, b)$$

These quantities can be used to obtain the chi-squared test statistic. The p-value p for this test can be obtained using the chi-squared distribution with $g-1$ degrees of freedom. Assume that the current patient is in category h, with $1 \leq h \leq g$. Then the voting law for the covariate is:

- Vote for A if $p < p^*$ and $E_{ha} > n_{ha}$
- Vote for B if $p < p^*$ and $E_{hb} > n_{hb}$
- Neutral otherwise

Suppose that the covariate is categorical (with a large amount of categories) : Most notably the clinical center where the patient receives treatment is a covariate that can have many categories. In this case, the procedure previously described may be ineffective for balancing categories with few patients. Therefore, the imbalance for a category is defined as the difference between the proportion of patients from group A in this category and $1/2$. To do so, a one sample binomial test is used.

Let's again assume that the current patient is in category h. Define n_h and n respectively the total amount of patients in the category h previously randomized and the total number of patients previously randomized . If $n_h \geq 20$ the normal approximation is used, and the law of the following test statistic is a standard gaussian :

$$z = \frac{\left(\frac{n_{ha}}{n_h} - \frac{n_a}{n}\right)}{\sqrt{\frac{n_a}{n}.\frac{n_b}{n}.\frac{1}{n_h}}}$$

From there, the p-value of the test can easily be computed. If $n_h < 20$, the exact two tailed p-value is used (see *Nonparametric Statistical Methods* by Hollander,

M. and Wolfe, D.A. [12]). In both cases, the voting law for the covariate is :

- Vote for A if $p < p^*$ and $n_{ha}/n_h < 1/2$
- Vote for B if $p < p^*$ and $n_{ha}/n_h > 1/2$
- Neutral otherwise

2.1.3 Minimization

Let's assume that m categorical covariates have to be balanced during the randomization process. For $1 \leq i \leq m$, let g_i be the number of categories for the covariate i.

The randomization procedures is the following :

1. The first patient is randomized in the treatment group A with probability $1/2$.
2. Assume that l-th patient randomized has the covariate profile $j_l = (j_1, ..., j_m)$. For $1 \leq i \leq m$, define $D_l^{(a)}(j_i)$ as the absolute difference in the number of patients in category j_i between treatment groups A and B if the patient were to go to group A. Let $D_l^{(b)}(j_i)$ be the same statistic for the treatment group B. For example, suppose that patient number 100 has category j_1 and that there are 9 patients from group A and 12 from group B with this category. Then $D_{100}^{(a)}(j_1) = |10 - 12| = 2$ and $D_{100}^{(b)}(j_1) = |13 - 9| = 4$.
3. The imbalance is measured with :

$$Imb_l^{(k)} = \sum_{i=m}^{I} [D_l^{(k)}(j_i)]^2, k = a, b;$$

The decision procedure for patient j is as follows :

- If $Imb_l^{(a)} < Imb_l^{(b)}$, treatment A is given with probability ξ.
- If $Imb_l^{(a)} > Imb_l^{(b)}$, treatment B is given with probability ξ.
- If $Imb_l^{(a)} = Imb_l^{(b)}$, treatment A is given with probability $1/2$.

Once again, ξ is called the biased coin probability.

Stratified permuted blocks and minimization were simulated using the *carat* package of *R*. The minimal sufficient balance randomization procedure was implemented from scratch. The implementation was successful, as the results obtained were similar to those observed in the original publication.

2.2 Simulating data

2.2.1 Generalities on NORTA

The NORTA method, described by Cario and Nelson [4], aims at simulating a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ with given marginal distributions and correlation matrix, noted Σ_X. To do so, a multivariate gaussian $\mathbf{Z} = (Z_1, \ldots, Z_d)$ is generated with specified correlation matrix Σ_Z. The univariate normal distribution function is then applied to each coordinate of $\mathbf{Z}$ to obtain a correlated vector of uniform marginals, noted $\mathbf{U}$. Finally, $\mathbf{X}$ is obtained by applying the corresponding inverse distribution function, also called quantile function, to each coordinate of U. Thus :

$$\mathbf{X} = \left(F_1^{-1}\left[\Phi\left(Z_1\right)\right], \ldots, F_d^{-1}\left(\left[\Phi\left(Z_d\right)\right]\right)\right)$$

Where

- $\mathbf{Z} \sim \mathcal{N}(0, 1)$ with correlation matrix Σ_Z
- Φ is the distribution function of a normal distribution with mean zero and variance one.
- $F_i^{-1}(u) = \inf\{x : F_i(x) \geq u\}$ is the quantile function of the desired marginal distribution.

Therefore, the goal is to find a correlation matrix for $\mathbf{Z}$ in order to obtain, on average, the desired correlation matrix for $\mathbf{X}$. In this article, both continuous and discrete (ordinal) variables were considered. In this case, the literature emphasis the use of the rank correlation instead of Pearson's linear correlation to specify the desired coefficients of Σ_X.
For $1 \leq i, j \leq d$, the rank correlation, which is analogous to Spearman's correlation, between X_i and X_j is defined by

$$\begin{aligned} r_{i,j}^X &= \mathrm{Corr}\left(F_i\left(X_i\right), F_j\left(X_j\right)\right) \\ &= \mathrm{Corr}\left(F_i \circ F_i^{-1} \circ \Phi\left(Z_i\right), F_j \circ F_j^{-1} \circ \Phi\left(Z_j\right)\right) \end{aligned}$$

where Corr designates Pearson's correlation. Define $\rho_{i,j} = \mathrm{Corr}\left(Z_i, Z_j\right)$, the coefficient located at the j-th column of the i-th row of Σ_Z. The NORTA rank correlation problem consists in finding $\rho_{i,j}^*$ such that, for $1 \leq i, j \leq d$, $r_{i,j}^X = \widetilde{r_{i,j}^X}$, where $\widetilde{r_{i,j}^X}$ is the desired rank correlation coefficient. This means that $\frac{d(d-1)}{2}$ equations have to be solved, since Σ_X is symmetrical.

2.2.2 Continuous – continuous case

Let $(i,j) \in [\![1:d]\!]$. If both X_i and X_j are continuous, then $F_i \circ F_i^{-1} = F_j \circ F_j^{-1} = I$, where I is the identity function. Therefore :

$$\begin{aligned} r_{i,j}^X &= \mathrm{Corr}\left(\Phi\left(Z_1\right), \Phi\left(Z_2\right)\right) \\ &= (6/\pi)\arcsin(\rho_{i,j}/2) \end{aligned}$$

This is a well known result, established by Pearson, Karl in 1907 [17]. As a result, the NORTA rank correlation problem is simply solved in this case by taking :

$$\rho_{i,j}^* = 2\sin(\pi \widetilde{r_{i,j}^X}/6)$$

2.2.3 Other cases

Let $(i,j) \in [\![1:d]\!]$. The rank correlation between X_i and X_j can be written as :

$$r_{i,j}^X = \frac{\mathbb{E}\left[F_i\left(F_i^{-1}\left(\Phi\left(Z_i\right)\right)\right) F_j\left(F_j^{-1}\left(\Phi\left(Z_j\right)\right)\right)\right] - \mu_{F_i}\mu_{F_j}}{\sigma_{F_i}\sigma_{F_j}}$$

where μ_{F_i} and μ_{F_j} are the means of the distribution functions of X_i and X_j, and σ_{F_i} and σ_{F_j} are the standard errors of X_i and X_j. These values can easily be integrated numerically. For example, the mean of F_i can be written as

$$\mu_{F_i} = \int_{-\infty}^{\infty} F_i(F_i^{-1}(\Phi(z)))\phi(z)dz$$

with $\phi(z)$ the density of a standard gaussian.
Let's define

$$g_{i,j}(\rho) = \mathbb{E}\left[F_i\left(X_i\right) F_j\left(X_j\right)\right]$$

such that

$$r_{i,j}^X = \frac{g_{i,j}(\rho) - \mu_{F_i}\mu_{F_j}}{\sigma_{F_i}\sigma_{F_j}}$$

The rank correlation problem is equivalent to finding $\rho_{i,j}^*$ such that $r_{i,j}^X - \widetilde{r_{i,j}^X} = 0$, which means finding a 0 (also called root) of the function $f_{i,j}$, defined as :

$$f_{i,j}(\rho) = g_{i,j}(\rho) - \mu_{F_i}\mu_{F_j} - \widetilde{r_{i,j}^X}\sigma_{F_i}\sigma_{F_j}$$

In order to evaluate f, an expression for $g_{i,j}$ has to be derived. This expression depends on whether both X_i and X_j are discrete, or if one of them is continuous.

Let's assume in a first place that one of the variable is not discrete.Without loss of generality, suppose that X_j has a discrete distribution over the integers and X_i has a continuous distribution. Then

$$\begin{aligned} g_{i,j}(\rho) = \sum_{l=-\infty}^{\infty} f_{j,l} \Bigg[\int_0^1 u_i \Bigg(\Phi\left(\frac{z_{j,l} - \rho\Phi^{-1}(u_i)}{\sqrt{1-\rho^2}}\right) \\ - \Phi\left(\frac{z_{j,l-1} - \rho\Phi^{-1}(u_i)}{\sqrt{1-\rho^2}} d\right)\Bigg) du_i \Bigg] \end{aligned}$$

where $f_{j,l} = F_j(l)$ and $z_{j,l} = \Phi^{-1}(f_{j,l})$, as shown by Channouf, Nabil, and Pierre L'Ecuyer. in 2009 [5],

Suppose now that both X_i and X_j are discrete variables. Then:

$$g_{i,j}(\rho) = \sum_{k=-\infty}^{\infty} p_{i,k+1} \sum_{l=-\infty}^{\infty} p_{j,l+1} \bar{\Phi}_\rho\left(z_{i,k}, z_{j,l}\right)$$

where $p_{u,v} = \mathbb{P}(X_u = v)$ if $u \in [\![1:d]\!]$ and v is an integer, and $\bar{\Phi}_\rho(x,y) = \int_x^\infty \int_y^\infty \phi_\rho\left(z_1, z_2\right) dz_1 dz_2$ with ϕ_ρ the bivariate normal density with covariance ρ (Avramidis, Athanassios N., Nabil Channouf and Pierre L'Ecuyer [1]).

In both cases, the following properties are true:

- f is continuous and strictly increasing on $[-1,1]$
- $\rho_{i,j}^* \in [-1,0]$ if $\widetilde{r_{i,j}^X}$ is negative and $\rho_{i,j}^* \in [0,1]$ if $\widetilde{r_{i,j}^X}$ is positive

Therefore, if say $\widetilde{r_{i,j}^X}$ is negative, then $f(-1)$ and $f(0)$ have different signs and a root of f belongs to the interval $[-1,0]$. Therefore, prerequisites to use the procedure *zero* of Brent [3], on the interval $[-1,0]$, are satisfied. This procedure has guaranteed convergence, and stops once it finds a number that is closer than ϵ to the true root. ϵ was chosen to be equal to 10^{-6}. If $\widetilde{r_{i,j}^X}$ is positive, the root finding procedure is the same, except Brent's method is applied on the interval $[0,1]$.

If the variables were continuous, their empirical quantiles and distribution functions were used. If the variables were binary, the quantiles and distribution functions of a Bernoulli with the empirical mean as parameter were used. The bivariate normal integral is challenging to evaluate. We have chosen to compute it using *algorithm 462* by Donelly [8], as recommended by Avramidis, Athanassios N., Nabil Channouf and Pierre L'Ecuyer, 2009 [1]. All the other numerical integrations were done using R. Piessens and E. deDoncker–Kapenga's Quadpack routines [18], with an absolute accuracy of 10^{-15}. Simulations were carried out to verify that the method had been implemented correctly. To do this, for given ρ parameters, large samples of $\mathbf{X}$ were calculated. The rank correlation matrix of this vector was then calculated. This provides consistent estimators of r^X, an idea explored in more details by Chen, Huifen. [6]. These estimators were extremely close to the values obtained via the NORTA implementation.

2.2.4 Addition of noise

Let $\Sigma_{Z^*} = (\rho_{i,j}^*)_{(i,j)\in[\![1:d]\!]}$. The estimated correlation matrix from real data may not be a perfect representation of the true underlying correlation structure. Thus,

in order to add variability to the correlation structures of the simulated datasets, noise is added to Σ_{Z^*}. For each simulation, instead of using Σ_{Z^*} to generate $\mathbf{Z}$, the matrix :

$$\Sigma_{Z^{**}} = \Sigma_{Z^*} + \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \dots & \epsilon_{1d} \\ \epsilon_{21} & \epsilon_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \epsilon_{d1} & \dots & \dots & \epsilon_{dd} \end{pmatrix}$$

is used, where for $1 \leq i, j \leq d$, $\epsilon_{ij} \sim \mathcal{N}(0, 0.025)$. This way, 95% of the coefficients of Σ_{Z^*} only vary by $\pm$ 0.05. Since the function $\rho \to g_{i,j}(\rho)$ is linear when $\rho \in [\![-0.95 : 0.95]\!]$, this also generally means that the coefficients of Σ_X vary by the same amount in comparison to the observed correlations in the original dataset.

In some cases, $\Sigma_{Z^{**}}$ is not positive definite and therefore not a suitable covariance matrix. In this study, if the matrix is not positive definite, another sample of ϵ is simply generated, and this step is repeated until $\Sigma_{Z^{**}}$ satisfies the conditions for being a covariance matrix.

2.2.5 Augmenting the Sepsiscool-I dataset

Sepsiscool-I is a multicenter, randomized controlled trial of 200 patients conducted in 2011 to assess whether fever control by external cooling reduces the need for vasopressors in septic shock. The trial was a success. However, the statistical power was insufficient to determine whether fever control by external cooling also helps reduce mortality.

It was hypothesized that the treatment was more effective among patients with lactate levels above 2 mmol/l. These patients' characteristics for the covariates that appeared to be the biggest predictors of mortality can be found in table 2.

	Whole sample (n=104)	Cooling (n=53)	No Cooling (n=51)
Age, yr (continuous)	60 (14) (22;86)	61 (15) (24;86)	59 (14) (22;86)
Vasopressor dose at baseline, mg/kg/min (continuous)	1 (0.87) (0.1;4.5)	1.3 (1) (0.1;4.5)	0.7 (0.5) (0.1;2.6)
FaO2/FIO2, mm Hg (continuous)	164 (81) (49;389)	162 (79) (49;360)	166 (85) (69;389)
Serum lactate level, mmol/L (continuous)	4.1 (2.4) (2.1;13)	4.2 (2.4) (2.1;13)	4.1 (2.4) (2.1;13)
Immunosuppression (binary)	22 (27%)	9 (26%)	13 (27%)
Acute respiratory distress syndrome (ARDS) (binary)	52 (50%)	24 (45%)	28 (56%)
Clinical center of randomization (categorical)			
Center 1	19 (18.3 %)	8 (15.0 %)	11 (21.6%)
Center 3	5 (4.8 %)	3 (5.7 %)	2 (3.92 %)
Center 4	3 (2.9 %)	1 (2.0 %)	2 (3.92 %)
Center 5	46 (44.2 %)	24 (45.3 %)	22 (43.14 %)
Center 7	12 (11.5 %)	6 (11.3 %)	6 (11.8 %)
Center 8	4 (3.9 %)	3 (5.7 %)	1 (2.0 %)
Center 9	15 (14.4 %)	8 (15.1 %)	7 (13.7 %)

Table 2: Patients from the Sepsiscool-1 trial with lactate levels above 2 mmol/l.
For continuous covariates : mean (standard error) (min;max)
For binary and categorical covariates : sample size of the category (percentage)

Only the patients with lactate levels above 2 mmol/l will be selected for the Sepsiscool-2 clinical trial, designed to assess whether fever control by external cooling significantly reduces mortality. Minimal Sufficient Balance will be used to balance over all the previous covariates, as well as the covariate for clinical center.

The full randomization protocol of this trial can be found in figure 3.

The dataset of the eligible patients from Sepsiscool-1, which will be called the original datatset, has a rank correlation matrix for the covariates where rank correlation can be calculated (binary and continuous ones) that can be found in figure 8.

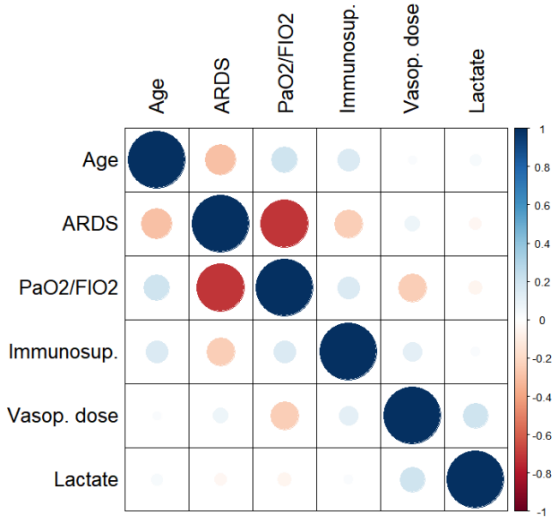


Figure 1: Rank correlation matrix of the Sepsicool 1 dataset (patients with lactate level above 2 mmol/l

In theory, the NORTA method could be applied to the entire correlation matrix. However, some associations could be due solely to statistical fluctuations. If they were included in the simulation, these fluctuations would be reproduced on all data sets. Therefore, rank correlation tests (see *Nonparametric Statistical Methods* by Hollander, M. and Wolfe, D.A. [13]) were performed between all covariates in pairs. Only associations with a test p-value below 0.3 were retained. This threshold is used in the Sepsiscool-2 protocol, and is consistent here as it delimits a gap between test p-values, as can be seen in figure 2

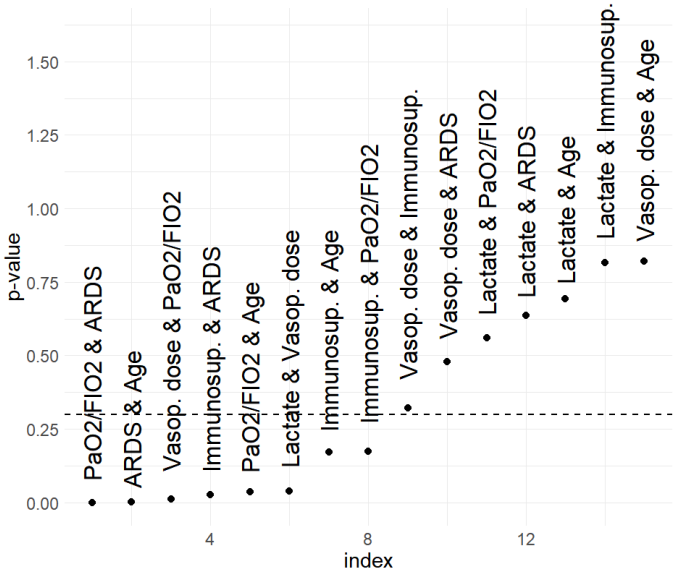


Figure 2: p-values of rank correlation tests between covariate pairs on the original dataset. Every association below the vertical line is kept in the NORTA procedure.

Since the correlation between the clinical center for randomization and the other variables can not be easily expressed, this covariate was generated independently from the others. It was simulated using a multinomial law so that the observed proportion of patients from each center was conserved.

Now that the associations kept are determined, the size of the simulated datasets have to be choosen. The Sepsiscool-2 protocol requires 820 patients, which is far more than the number of Sepsiscool-1 patients who would have been eligible to participate in Sepsiscool-2 (104). The number 820 is obtained because the authors of the Sepsicool-2 protocol assume that the mortality will be 42% among patients that receive the treatment and 52% among patient that don't, and they want 80% power and a significance level of 5% to show the treatment's effectiveness. If we stick to the data from the original dataset, the mortality is 45% among patients that receive external cooling and 62% among patients that don't. Using these values and the sample size calculation described by Chow, Shein-Chung, et al. in *Sample size calculations in clinical research* [7], we deduct that for 80% power and a 5% significance level, 270 patients would be necessary to prove that cooling significantly reduces mortality. This number can be rounded to 300 as additionnal patients are always included. Both these sample sizes were selected because they allow to test the randomization algorithms in different conditions.

5000 datasets were generated for both scenarios using NORTA. On top of that, 5000 uncorrelated datasets were generated for both scenarios by independently bootstraping the covariates from the original dataset. These uncorrelated datasets served as control to study the impact of the covariates' correlation structures on the randomization procedures.

Wilcoxon signed-rank tests were performed between each generated covariate and the equivalent covariate from the original dataset. Generated datasets were not retained if the p-value of one of the tests was less than 10^-4. This threshold is common in genomics for example, and is low because of the multiplicity of tests.

Each randomization algorithm was then applied once to each generated dataset. Thus, 4 different treatment allocations were obtained for each dataset. The parameters for each randomization procedure were the following :

For minimal sufficient balance, the Sepsiscool-2 protocol indicates a biased coin probability ξ of 0.7 and an imbalance threshold p^* of 0.3. Those were kept for the simulations. Moreover, the covariate for clinical center was to considered to be a covariate with a large amount of categories.

It is necessary to discretize the quantitative covariates when using minimization. Thus a parameter for minimization is the number of same size class in which the quantitative variables are divided. Two different procedures were choosen, one where the four continuous covariates where divided in two classes and one where they were divided in four classes. In both these randomization procedures, the biased coin probability parameter for Minimization was also chosen to be 0.7. It is important to keep in mind that this choice does not guarantee the same number of non-random allocations as in Minimal Sufficient Balance.

Finally, block size in stratified permuted randomization was set to four, which is one of the more popular sizes. All the procedures were stratified on ARDS, except the stratified permuted blocks which were stratified on ARDS and clinical center. Table 3 summarises the choices made for the simulations.

Scenarios	Procedures	Stratification	Parameters
Scenario 1 : correlated datasets, 300 patients	Minimial Sufficient Balance	ARDS	$\xi = 0.7$ $p^* = 0.3$
Scenario 2 : uncorrelated datasets, 300 patients	Minimization	ARDS	$\xi = 0.7$ #Class = 2
Scenario 3 : correlated datasets, 820 patients	Minimization	ARDS	$\xi = 0.7$ #Class = 4
Scenario 4 : uncorrelated datasets, 820 patients	Stratified Permuted Blocks	ARDS & Clinical Center	Block size = 4

Table 3: Summary of the parameters used in each simulation

2.2.6 Simulating the outcome

Applying the randomization algorithms allows to obtain an allocation of treatment for each dataset. In order to carry out a more in-depth analysis of the effectiveness of the different algorithms, the outcome (mortality) had to be simulated. For this simulation, it was hypothesized that the cooling treatment was effective in reducing mortality. The effectiveness of the treatment was calculated on the original dataset. To do so, a binomial logistic regression model with logit as the link function and without intercept was fitted on the predictors to explain mortality. Not using an intercept allowed to obtain a coefficient for each category of the center variable. The reference categories were 0 for the binary covariates. The coefficients obtained for the predictors can be found in section B.2. The coefficient for the adjusted treatment effect was equal to -9.5×10^{-2}, with standard error 5.7×10^{-1}.

For each dataset, a treatment coefficient, $\beta_{treatment}$, was calculated by generating an observation of the variable $|\mathbf{T}|$, where $\mathbf{T} \sim \mathcal{N}(-9.5 \times 10^{-2}, 5.7 \times 10^{-1})$. This way, treatment effect was guaranteed to be positive and uncertainty was taken into account. Let's define $\beta_{cov} = (\beta_{age}, ..., \beta_{center9})$. For a given simulated patient from a generated dataset, define Y a Bernoulli variable that indicates the outcome ($Y = 1$ in case of death, $Y = 0$ in case of survival). Moreover, define X_{cov} and $X_{treatment}$ respectively the vector that indicates the patient's covariates values and the treatment he received. By considering a logistic regression in reverse, the dependence between Y and the patient's characteristics can be modeled by :

$$\ln(\tfrac{\mathbb{P}(Y=1)}{1-\mathbb{P}(Y=1)}) = \beta_{cov}X_{cov} + \beta_{treatment}X_{treatment}.$$

Therefore, the outcome can be simulated by generating an observation of Y, with

$$\mathbb{P}(Y = 1) = logit^{-1}(\beta_{cov}X_{cov} + \beta_{treatment}X_{treatment})$$

2.3 Evaluation criteria for randomization algorithms

2.3.1 Root Mean Squared Error

The Root Mean Squared Error (RMSE) indicates the extent to which the randomization procedure recovers the treatment coefficient with which the outcome was simulated. For a given randomization procedure and scenario, let $k \in [\![1 : 5000]\!]$ index the datasets. Let $\beta^k_{estimated}$ be the treatment coefficient of a logistic regression model done on dataset number k, with mortality as the explanatory variable and adjusted on the covariates. RMSE can be written as :

$$\text{RMSE} = \textstyle\sum_{k=1}^{n}(\beta^k_{estimated} - \beta^k_{treatment})^2, n = 5000$$

If the RMSE is too high, the randomization doesn't permit to recover $\beta_{treatment}$. In this case, something is wrong with the procedure, and no further analysis can be conducted. The RMSE serves as a last verification that every procedure is working well. The Mean Squared Error (MSE) is obtained by squaring the RMSE. It has the same interpretation and is also a very common indicator.

2.3.2 Imbalance tests

Imbalance tests were used to quantify the similarity of patient characteristics in the two treatment groups. Each randomization procedure was stratified on ARDS, as the treatment subgroups must also be balanced in the Sepsiscool-2 protocol. Therefore, global and stratified imbalance tests were performed.

For categorical covariates, chi-squared tests were performed. For quantitative covariates, Wilcoxon signed-rank tests were performed, as distributions were generally not normal in simulated datasets. Student's t-tests were also performed because of how common they are and because they correspond to what the MSB algorithm is designed to balance.

If a test's p-value was below 5%, it was considered that the groups were not balanced. This way of quantifying imbalance is the most common in post-hoc analysis of clinical trials. The algorithms that resulted in the least amount of imbalance on average were preferred.

2.3.3 Adjusted and unadjusted power

For given datasets and randomization procedures, adjusted and unadjusted power are calculated through a statistical test evaluating odds ratio equality : let p_a be the proportion of patients who died and received the treatment (fever control) and p_b be the proportion of patients who died in the control group. The Odd Ratio (OR) between the two group is defined as :

$$\text{OR} = \frac{p_a(1-p_a)}{p_b(1-p_b)}$$

The hypothesis are the following :

$$\text{H0} : \text{OR} = 1$$

$$H1 : OR \neq 1$$

Let $\kappa = \frac{n_a}{n_b}$ be the size ratio between the two treatment groups. The unadjusted power can be obtained with the following formula :

$$1 - \beta = \Phi\left(z - z_{1-\alpha/2}\right) + \Phi\left(-z - z_{1-\alpha/2}\right)$$

with

- $z = \frac{\ln(OR)\sqrt{n_b}}{\sqrt{\frac{1}{\kappa p_a(1-p_a)} + \frac{1}{p_b(1-p_b)}}}$
- α the desired Type 1 error (5% here)
- $z_{1-\alpha/2}$ the quantile $1 - \alpha/2$ of a standard Gaussian.

Adjusted power can be obtained by replacing the OR in the formula of z by the adjusted OR. The adjusted OR is equal to $\exp(\beta_{estimated})$. 95% confidence intervals were obtained for the mean adjusted and unadjusted power using the central limit theorem. The algorithms with the best mean adjusted and unadjusted power were preferred.

2.3.4 Signifiance

Global and stratified likelihood ratio tests with Type 1 errors of 5% were conducted to asses the proportion of treatment effect that would be considered signifiative. If the tests were significative on both subgroups (ARDS and No ARDS), the closed-testing procedure was considered significative. It is an important aspect of the Sepsiscool-2 protocol, as the treatment would ideally be efficient in both cases. The randomization procedures that displayed the highest level of signifiance were preferred.

All the analysis and simulations were done using R version 4.2.3.

3 Results

The following section presents the results for the two scenarios in which correlated datasets were used. For ease of reading, the results tables for scenarios 2 and 4, where uncorrelated datasets were used, can be found in the appendix (sections B.3 and B.4).

3.1 Scenarios 1 & 2 : 300 patients

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
MSE	0.10	0.10	0.10	0.10
RMSE	0.31	0.32	0.31	0.32

Table 4: Mean Squared Error and Root Mean Squared Error calculated for each randomization algorithm on 5000 generated datasets of size 300.

All algorithms have similar results for the MSE and RMSE criteria, whether the datasets are correlated or not (see table 12). No algorithm should be excluded due to poor results, allowing the analysis to continue.

	MSB		Blocks (4)		Minimization (2 classes)		Minimization (4 classes)	
	No ARDS	ARDS	No ARDS	ARDS	No ARDS	ARDS	No ARDS	ARDS
Centers	0.14	0.18	0.00	0.00	0.34	0.42	0.36	0.26
Immunosupression	0.00	0.02	3.56	2.60	0.02	0.12	0.00	0.06
Age	0.02 (0.00)	0.00 (0.00)	5.08 (5.34)	4.80 (4.98)	0.32 (0.66)	0.68 (1.28)	0.38 (0.54)	0.50 (0.64)
Vasopressor dose	0.24 (0.02)	0.26 (0.00)	4.96 (5.20)	4.86 (4.9)	0.18 (0.76)	0.20 (1.02)	0.18 (0.50)	0.32 (0.56)
PaO2/FIO2	0.04 (0.04)	0.14 (0.02)	4.96 (4.98)	5.46 (5.56)	1.78 (2.22)	1.66 (0.88)	0.44 (0.90)	0.38 (0.70)
Serum lactate	0.82 (0.06)	0.78 (0.02)	4.98 (4.76)	4.86 (5.14)	0.32 (1.8)	0.44 (1.98)	0.44 (0.96)	0.30 (1.24)

Table 5: Percentage of p-value of stratified imbalance tests below 5% on the 5000 generated datasets of size 300. The imbalance tests are Wilcoxon signed-rank test and Student's t-test (between parenthesis) for quantitative covariates and chi-squared test for qualitative covariates.

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
ARDS	0.92	0.00	0.00	0.00
Centers	0.32	0.00	0.62	0.54
Immunosuppression	0.08	2.96	0.00	0.00
Age	0.20 (0.10)	4.12 (3.98)	0.22 (0.84)	0.24 (0.24)
Vasopressor dose	0.16 (0.06)	4.92 (4.78)	0.24 (0.94)	0.20 (0.60)
PaO2/FIO2	0.60 (0.52)	0.70 (0.96)	0.02 (0.08)	0.00 (0.00)
Serum lactate	0.76 (0.10)	4.86 (4.60)	0.34 (2.04)	0.32 (1.24)

Table 6: Percentage of p-value of global imbalance tests below 5% on the 5000 generated datasets of size 300. The imbalance tests are Wilcoxon signed-rank test and Student's t-test (in parenthesis) for quantitative covariates and chi-squared test for qualitative covariates.

Firstly, for block randomization, the percentage of p-values below 0.05 for the stratified and global imbalance tests is around 5% for all covariates, with the exception of ARDS and centers. This is an expected result, given the definition of Type 1 error and the fact that the procedure was stratified on these last two covariates. Imbalances are on average less frequent if the covariates are correlated with ARDS, one of the stratified variables.

We note that the percentage of rejected tests is extremely low for the three remaining randomization algorithms. For stratified imbalance tests, the MSB algorithm outperforms the others on all covariates for Student's t-tests and chi-squared tests. When Wilcoxon signed-rank tests are used, MSB also outperforms both Minimization procedures on all covariates, with the exception of vasopressor dose at randomization and serum lactate level.

For global imbalance tests, Minimization is better than MSB for balancing ARDS, the covariates on which the adaptive procedures were stratified, but slightly worse for balancing centers. It also equals or surpasses MSB on all the Wilcoxon signed-rank tests. However, MSB still performs better when Student's t-tests are used.

Overall, for 300 patients, Minimal Sufficient Balance is the best for subgroup balance. Its results are slightly inferior to those of Minimization when using non-stratified Wilcoxon signed-rank tests, and slightly superior when using non-stratified Student's t-tests. Minimization performs better if quantitative covariates are divided into 4 classes.

For uncorrelated datasets, stratified permuted blocks perform worse, as the percentage of rejected test is always around 5%. However, Minimal Sufficient Balance and Minimization perform similarly (see table 13 and table 14).

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
Mean unadjusted power	35.3 (34.5;36.1)	34.2 (33.4;35.0)	35.0 (34.1;35.8)	35.4 (34.6;36.2)
Mean adjusted power	53.0 (52.0;54.0)	52.4 (51.4;53.4)	52.1 (51.2;53.1)	52.6 (51.6;53.6)
% of significant global treatment effect	38.6	37.7	37.5	38.9
% of significant treatment effect (ARDS)	25.1	25.5	24.9	26.0
% of significant treatment effect (No ARDS)	26.9	25.7	26.2	26.2
% of significant closed testings	11.6	11.4	11.7	12.3

Table 7: Ability of randomization procedures to show treatment's efficacy, 5000 correlated datasets of size 300.
Between parenthesis : 95% confidence interval calculated using the central limit theorem

The confidence intervals for adjusted and unadjusted power can be visualised with figure 4 and figure 5 in section B.5. Minimal Sufficient Balance and 4 classes Minimization have the best mean unadjusted and adjusted power. However, these results are not statistically significant. These two procedures are also the best for obtaining statistically significant treatment effects, especially Minimization which slightly surpasses MSB on all the significance criteria except the percentage of significant treatment effect on patients withouth ARDS. For uncorrelated datasets, the results for power are the same, and MSB surpasses 4 classes Minimization in all significance criteria except in the percentage of significant global treatment effect (see table 15).

3.2 Scenarios 3 & 4 : 820 patients

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
MSE	0.18	0.18	0.18	0.18
RMSE	0.03	0.03	0.03	0.03

Table 8: Mean Squared Error and Root Mean Squared Error calculated for each randomization algorithm on 5000 generated datasets of size 820.

Once again, all the algorithms have similar results for the MSE and RMSE criteria whether the datasets are correlated or not (see table 16). No algorithm has to be excluded because of poor results, further analysis can be pursued.

	MSB		Blocks (4)		Minimization (2 classes)		Minimization (4 classes)	
	No ARDS	ARDS	No ARDS	ARDS	No ARDS	ARDS	No ARDS	ARDS
Centers	0.00	0.02	0.00	0.00	0.02	0.00	0.06	0.04
Immunosuppression	0.00	0.00	3.38	4.10	0.00	0.02	0.00	0.00
Age	0.00 (0.00)	0.00 (0.00)	4.80 (4.86)	4.94 (5.08)	0.12 (0.44)	0.16 (0.56)	0.00 (0.02)	0.04 (0.12)
Vasopressor dose	0.06 (0.00)	0.08 (0.00)	4.66 (4.68)	4.54 (4.80)	0.10 (0.60)	0.02 (0.64)	0.02 (0.06)	0.00 (0.08)
PaO2/FIO2	0.00 (0.00)	0.00 (0.00)	4.52 (4.52)	4.96 (4.66)	1.40 (1.60)	1.08 (0.22)	0.06 (0.22)	0.04 (0.10)
Serum lactate	0.68 (0.00)	0.70 (0.00)	4.84 (4.48)	4.98 (5.00)	0.02 (1.36)	0.04 (1.26)	0.00 (0.24)	0.00 (0.34)

Table 9: Percentage of p-value of stratified imbalance tests below 5% on the 5000 generated datasets of size 820. The imbalance tests are Wilcoxon signed-rank test and Student's t-test (in parenthesis) for quantitative covariates and chi-squared test for qualitative covariates.

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
ARDS	0.44	0.00	0.00	0.00
Centers	0.02	0.00	0.00	0.04
Immunosuppression	0.00	3.90	0.00	0.00
Age	0.04 (0.02)	3.74 (3.76)	0.02 (0.34)	0.00 (0.06)
Vasopressor dose	0.16 (0.00)	4.94 (4.74)	0.02 (0.44)	0.00 (0.10)
PaO2/FIO2	0.28 (0.18)	0.46 (0.82)	0.00 (0.04)	0.00 (0.00)
Serum lactate	0.66 (0.00)	4.92 (4.46)	0.06 (1.38)	0.04 (0.36)

Table 10: Percentage of p-value of global imbalance tests below 5% on the 5000 generated datasets of size 820. The imbalance tests are Wilcoxon signed-rank test and Student's t-test (in parenthesis) for quantitative covariates and chi-squared test for qualitative covariates.

For block randomization, the percentage of p-values below 0.05 for stratified and global imbalance tests is always around 5% for all covariates, with the exception of ARDS and centers.

After 820 patients are randomized, the percentage of rejected test is almost null for all covariate adaptative randomization procedures. For MSB, no signficant imbalances were observed for Student's t tests and chi-squared tests out of the 5000 simulations. For Wilcoxon signed-rank tests, MSB surpasses or equals 2 and 4 classes Minimization for all covariates except serum lactate level.

For global imbalance tests, Minimization is better than MSB for balancing ARDS, the covariates on which the adaptive procedures were stratified, but slightly worse for balancing centers. It also equals or surpasses MSB on all the Wilcoxon signed-rank tests. However, MSB still performs better when Student's t-tests are used. The best performances for Minimization are obtained when the continuous variables are

divided in 4 classes.

Overall for 820 patients, the conclusions are the same than for 300 patients for both correlated and uncorrelated datasets, except the imbalance are even less common when adaptive randomization algorithms are used.

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
Mean unadjusted power	51.1 (50.1;52.1)	50.6 (49.6;51.6)	51.9 (51.0;52.9)	50.8 (49.8;51.9)
Mean adjusted power	65.9 (64.9;66.9)	65.7 (64.6;66.7)	66.7 (65.6;67.7)	66.0 (65.0,67.0)
% of significant global treatment effect	57.1	56.7	58.6	56.9
% of significant treatment effect (ARDS)	44.2	43.3	45.2	43.4
% of significant treatment effect (No ARDS)	44.7	44.2	45.1	44.3
% of significant closed testings	31.6	30.2	31.6	30.5

Table 11: Ability of randomization procedures to show treatment's efficacy, 5000 correlated datasets of size 820.
Between parenthesis : 95% confidence interval calculated using the central limit theorem

This time, 2 classes Minimization has the best mean adjusted and unadjusted power. However, the differences in power between the adaptative randomization procedures are not statistically significant (see figure 6 and figure 7).

2 classes Minimization also is the best for obtaining statistically significant treatment effects, followed by MSB and 4 classes Minimization. For uncorrelated datasets, 4 classes Minimization has better performance than the other procedures for power and showing treatment effects (see table 19).

4 Discussion

Overall, this study shows that, for clinical data, Minimization and Minimal Sufficient Balance are as good or better than Stratified Permuted Blocks to show treatement efficiency, and largely outperform this procedure when it comes to limiting imbalances. When using Minimization or Minimal Sufficient Balance, the probability of getting unusable data because of imbalance on an important covariate is extremely low.

Minimal Sufficient Balance displayed excellent balancing abilities with both Student's t-test and Wilcoxon signed-rank test, no matter the sample size. Minimization was only more efficient in the particular case where non-stratified Wilcoxon signed-rank tests were used. However, improvements to MSB have recently been suggested by Johns, Hannah, et al. [14], and this new version could outperform Minimization in every aspects.

In practice, covariate adaptative randomization procedures are rarely stratified, as the important covariates are included in the randomization. In this case MSB is simply better for balancing.

The Sepsiscool-2 trial has the particularity of having an interim analysis after half of the patients are randomized, after which the trial can be stopped for futility, continued on one of the subgroups (ARDS and no ARDS) or on both subgroups (see figure 3 for the full randomization procedure of Sepsiscool-2). Thus, the covariates have to be balanced on both subgroups, which justifies the stratification.

Even in such a situation, using Minimal Sufficient

Balance is still perfectly reasonable as it is much better than Minimization at preserving allocation randomness (S. D. Lauzon et al. [16]) and still guarantees extremely low chances of imbalance, especially if the patient's covariates are expected to be normally distributed.

The best procedure at showing treatment efficiency seems to depend a lot on the data, as the results were different between the correlated and uncorrelated datasets. Therefore, there doesn't seem to be a right algorithm to maximise the chance of showing treatment efficiency. Further analysis could be done to try and find the optimal adaptative randomization algorithm depending on the expected statistical properties of a trial's dataset.

The strongest point of the study is the emphasis that was put on the generation of the data. The NORTA methods allows for a lot of flexibility in the choice of the associations to consider, the choice of the distribution function used and the possibility of adding noise to the correlation matrix. The Wilcoxon signed-rank test performed after completion of NORTA ensures data quality. Nearly no datasets were rejected, which means that the threshold of 10^-4 could have been lowered. Still, the simulated patients were all different and plausible. This method permits to simulate the algorithms on a large variety of datasets, instead of just applying them 5000 times on the same dataset. Thus, the fact that the Sepsiscool-2 dataset is not yet available has little impact on the methodology, as the goal of the study is not to guess the trial's result or critique its protocol, but rather to use it as an example of trial that features adaptative randomization.

However, the implementation of this method has a number of limitations. Firstly, the empirical distribution and quantile functions were used in the NORTA method for quantitative covariates. In this case, the equality $F \circ F^{-1} = I$ is in fact false. However, in this study, the difference was generally of the order of 10^{-3}, and it was therefore neglected because this was the best way of implementing the method.

A bigger problem comes from the way NORTA itself works, and the fact that non-ordinal categorical variables cannot be included in the correlation matrix. In order to take into account the specific correlation structure of each center, we could have calculated seven different correlation matrices, one for each center, and then applied the NORTA method separately to each matrix. However, we would have had to calculate three correlation matrices with fewer than 5 individuals. According to Bonett, Douglas G., and Thomas A. Wright [2] the standard error of the rank correlation coefficient is greater than 0.5 in this case, leaving far too much uncertainty.

Therefore, the hypothesis that the correlation structure of the patients was the same in each center was formulated, and the center covariate was simulated independentely from the other covariates. In order to asses how true this hypothesis was, we conducted a Mixed Data Factorial Analysis (Escofier, B. [10]), using the R package FactoMineR. The center variable was considered to be a supplementary variable. Missing data was imputed using MissMDA (Vincent Audigier, François Husson, Julie Josse [22]), by keeping the first four factorial axes. V-test, as described by Escofier, Brigitte, and Jérôme Pagès [11], were used to evaluate the association between each center and each factorial axe. For a given factorial axe and center, the test's null hypothesis is that the mean of the factor is the same between the patients from the center and the other patients. The interesting aspect of this test is that it takes into account sample sizes. If the test statistic's value belongs to the interval [-1.96,1.96], it is significative at a 5% level. Despite the multiplicity of tests, not a single test was significative as you can see in table 22. Thus, the hypothesis is fairly solid.

In addition to the noise added to the correlation matrix, it would have been conceivable to add Gaussian noise after the datasets had been generated. However, we chose not to do this as the addition of noise would bias downwards all the correlations observed on the generated datasets, and all the work done to find the appropriate correlation matrix for NORTA would have been for nothing. In addition, the covariates in the Sepsiscool-1 dataset are generally rounded to one decimal place, and this aspect was reproduced in the simulated dataset to better fit the original data. Adding continuous noise would not have made sense in this situation.

The outcome was calculated using the adjusted Odd Ratio. This ratio is way closer than 1 than the two OR used to compute the required sample sizes (0.90 versus 0.50 and 0.67) for 80% power and a Type-I error of 5%. The sample size was not calculated using the adjusted Odd Ratio, because 14547 patients would have been required, which is extremely unusual for a clinical trial in reanimation. This explains why the percentage of significant global treatment effect is below 80%. The adjusted and unadjusted treatment effect serve as upper and lower bounds for this quantity, which can be interpreted as an estimation of the real power. The Type-I error could not be evaluated because of the assumption that the treatment effect was positive.

Finally, the variable used to simulate the outcome

described whether patients died before leaving the hospital. The primary outcome of Sepsiscool-2 is mortality at day 60, which is not exactly the same thing. Once again, as the aim of this study was simply to follow the protocol as closely as possible and not to try to predict the outcome, this aspect has little impact on the results.

5 Conclusion

In conclusion, we have shown that Minimal Sufficient Balance is an extremely performant covariate adaptative randomization algorithm that minimizes the risk of imbalance while guaranteeing a high probability of showing the treatment's efficiency. Further studies need to be made to determine which randomization procedure is the most likely to show treatment's effectiveness when writing the protocol of a clinical trial. Minimization slightly outperforms Minimal Sufficient Balance in a very particular situation.

References

[1] Nabil Channouf Avramidis Athanassios N. and Pierre L'Ecuyer. "Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal-copula dependence." In: *INFORMS Journal on Computing* 21.1 (2009), pp. 88–106.

[2] Douglas G. Bonett and Thomas A. Wright. "Sample size requirements for estimating Pearson, Kendall and Spearman correlations." In: *Psychometrika* 65 (2000), pp. 23–28.

[3] Richard P. Brent. "An algorithm with guaranteed convergence for finding a zero of a function." In: *The computer journal* 14.4 (1971), pp. 422–425.

[4] Marne C. Cario and Barry L. Nelson. *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix.* Tech. rep. Evanston, Illinois: Department of Industrial Engineering and Management Sciences, Northwestern University, 1997.

[5] Nabil Channouf and Pierre L'Ecuyer. "Fitting a normal copula for a multivariate distribution with both discrete and continuous marginals." In: *Proceedings of the 2009 winter simulation conference (WSC)* (2009), pp. 352–358.

[6] Huifen. Chen. "Initialization for NORTA: Generation of random vectors with specified marginals and correlations." In: *INFORMS Journal on Computing* 13.4 (2001), pp. 312–331.

[7] Shein-Chung Chow and Yuliya Lokhnygina Jun Shao Hansheng Wang. *Sample size calculations in clinical research.* CRC press., 2017.

[8] Thomas G. Donnelly. "Algorithm 462: Bivariate normal distribution." In: *Communications of the ACM* 16.10 (1973), p. 638.

[9] Bradley. Efron. "Forcing a sequential experiment to be balanced." In: *Biometrika,* 58.3 (1971), pp. 403–417.

[10] B. Escofier. "Traitement simultané de variables qualitatives et quantitatives en analyse factorielle." In: *Les cahiers de l'analyse des données* 4.2 (1979), pp. 137–146.

[11] Brigitte Escofier and Jérôme Pagès. "Analyses factorielles simples et multiples". In: *Dunod, Paris* 284 (1998), pp. 49–51.

[12] M. Hollander and D.A. Wolfe. *Nonparametric Statistical Methods.* New York : John Wiley and Sons, 1973, pp. 15–22.

[13] M. Hollander and D.A. Wolfe. *Nonparametric Statistical Methods.* New York : John Wiley and Sons, 1973, pp. 185–194.

[14] Hannah Johns and Leonid Churilov Dominic Italiano Bruce Campbell. "Common scale minimal sufficient balance: An improved method for covariate-adaptive randomization based on the Wilcoxon-Mann-Whitney odds ratio statistic". In: *Statistics in Medicine* 41.10 (2022), pp. 1846–1861.

[15] S. D. Lauzon et al. "Impact of minimal sufficient balance, minimization, and stratified permuted blocks on bias and power in the estimation of treatment effect in sequential clinical trials with a binary endpoint." In: *INFORMS Journal on Computing* 31.1 (2022), pp. 184–204.

[16] S. D. Lauzon et al. "Statistical properties of minimal sufficient balance and minimization as methods for controlling baseline covariate imbalance at the design stage of sequential clinical trials". In: *Statistics in Medicine* 39.19 (2020), pp. 2506–2517.

[17] Karl. Pearson. *On further methods of determining correlation.* Vol. 16. Dulau and Company, 1907.

[18] Robert Piessens et al. *Quadpack: a subroutine package for automatic integration.* Vol. 1. Springer Science Business Media, 2012.

[19] Stuart J. Pocock and Richard Simon. "Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial." In: *Biometrics* (1975), pp. 103–115.

[20] William F. Rosenberger and John M. Lachin. *Randomization in clinical trials: theory and practice.* John Wiley Sons, 2015.

[21] Frédérique Schortgen et al. "Fever control using external cooling in septic shock: a randomized controlled trial". In: *American journal of respiratory and critical care medicine* 185.10 (2012), pp. 1088–1095.

[22] Julie Josse Vincent Audigier François Husson. "A principal component method to impute missing values for mixed data". In: *Advances in Data Analysis and Classification* 10 (2016), pp. 5–26.

[23] Bernard L. Welch. "The generalization of 'STUDENT'S'problem when several different population varlances are involved." In: *Biometrika* 34.1-2 (1947), pp. 28–35.

[24] Michael D. Hill Zhao Wenle and Yuko Palesch. "Minimal sufficient balance-a new strategy to balance baseline covariates and preserve randomness of treatment allocation". In: *Stat Methods Med Res.* 24.6 (2015), pp. 989–1002.

A Lists of figures and tables

List of figures

List of tables

B Appendix

B.1 Full randomization protocol of Sepsiscool-2

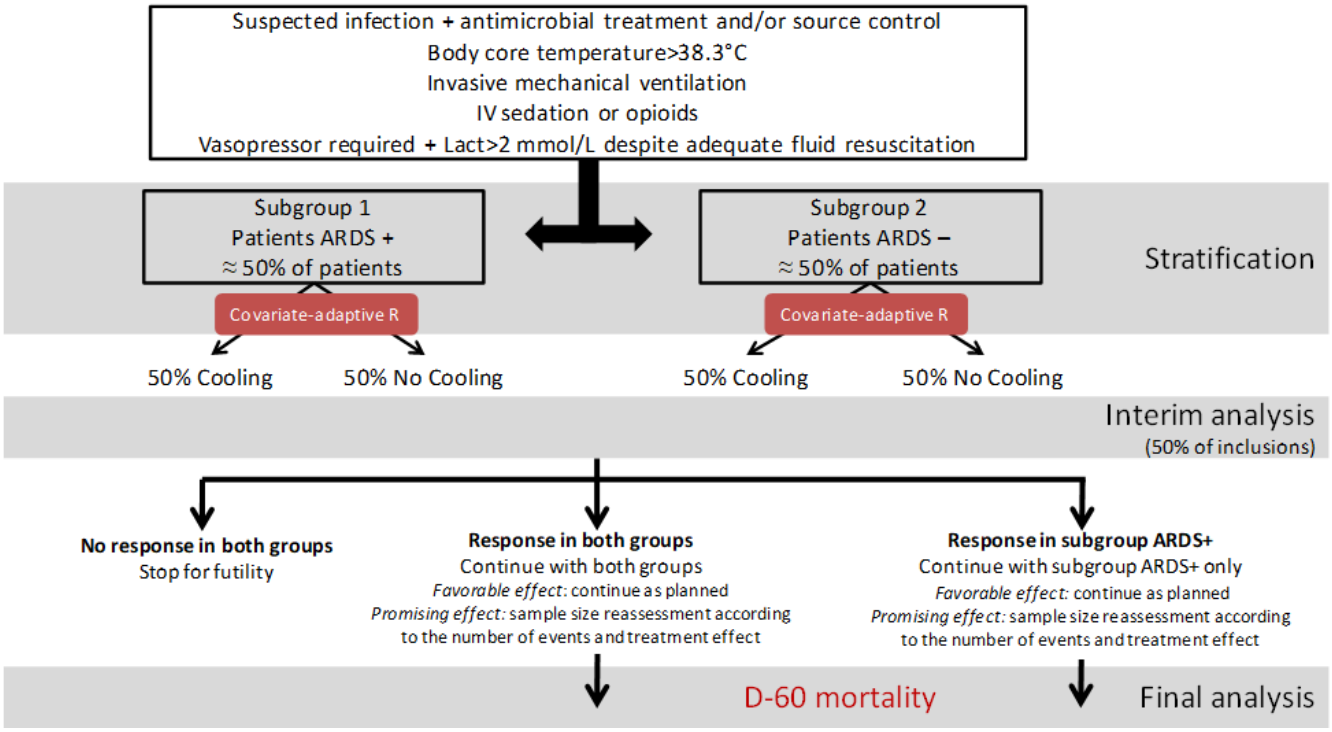


Figure 3: Full randomization protocol of the Sepsiscool-2 clinical trial. Source : Sepsiscool-2 research protocol

B.2 Parameters for simulating the outcome :

- $\beta_{age} = 4.0 \times 10^{-2}$
- $\beta_{ARDS} = 5.3 \times 10^{-1}$
- $\beta_{PF} = -4.3 \times 10^{-4}$
- $\beta_{immunosup} = 3.3 \times 10^{-1}$
- $\beta_{vasop} = 1.8 \times 10^{-1}$
- $\beta_{lactate} = 3.1 \times 10^{-1}$
- $\beta_{center1} = -3.2$
- $\beta_{center3} = -5.1$
- $\beta_{center4} = -20.7$
- $\beta_{center5} = -3.9$
- $\beta_{center7} = -4.2$
- $\beta_{center8} = 13.9$
- $\beta_{center9} = -5.0$
- $\beta_{treatment} = -9.5 \times 10^{-2}$

B.3 Scenario 1 : 300 patients (Uncorrelated datasets)

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
MSE	0.10	0.10	0.10	0.10
RMSE	0.32	0.32	0.31	0.32

Table 12: Mean Squared Error and Root Mean Squared Error calculated for each randomization algorithm on 5000 generated uncorrelated datasets of size 300.

	MSB		Blocks (4)		Minimization (2 classes)		Minimization (4 classes)	
	No ARDS	ARDS	No ARDS	ARDS	No ARDS	ARDS	No ARDS	ARDS
Centers	0.24	0.32	0.00	0.00	0.40	0.28	0.26	0.40
Immunosupression	0.00	0.02	3.06	2.78	0.02	0.04	0.04	0.00
Age	0.02 (0.02)	0.04 (0.02)	5.32 (5.24)	5.14 (5.74)	0.40 (0.92)	0.36 (1.00)	0.48 (0.58)	0.36 (0.58)
Vasopressor dose	0.28 (0.14)	0.20 (0.00)	5.36 (5.08)	4.70 (4.44)	0.26 (1.12)	0.28 (1.02)	0.48 (0.64)	0.32 (0.78)
PaO2/FIO2	0.04 (0.02)	0.04 (0.00)	4.98 (5.16)	4.74 (4.50)	0.50 (0.50)	0.30 (0.58)	0.30 (0.48)	0.34 (0.54)
Serum lactate	0.54 (0.04)	0.60 (0.02)	5.06 (5.18)	5.08 (4.70)	0.26 (1.78)	0.10 (1.84)	0.50 (1.24)	0.48 (1.14)

Table 13: Percentage of p-value of stratified imbalance tests below 5% on the 5000 generated uncorrelated datasets of size 300. The imbalance tests are Wilcoxon signed-rank test and Student's t-test (between parenthesis) for quantitative covariates and chi-squared test for qualitative covariates.

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
ARDS	0.76	0.00	0.00	0.00
Centers	0.42	0.00	0.48	0.32
Immunosuppression	0.02	3.54	0.04	0.00
Age	0.08 (0.02)	5.56 (5.64)	0.26 (0.80)	0.28 (0.48)
Vasopressor dose	0.40 (0.06)	5.06 (5.24)	0.26 (1.16)	0.30 (0.68)
PaO2/FIO2	0.06 (0.02)	5.00 (4.82)	0.34 (0.44)	0.24 (0.52)
Serum lactate	0.92 (0.04)	5.24 (4.88)	0.32 (1.86)	0.36 (1.66)

Table 14: Percentage of p-value of imbalance tests below 5% on the 5000 generated uncorrelated datasets of size 300. The imbalance tests are Wilcoxon signed-rank test and Student's t-test (in parenthesis) for quantitative covariates and chi-squared test for qualitative covariates.

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
Mean unadjusted power	35.0 (34.2;35.8)	34.2 (33.4;35.0)	34.7 (33.8;35.5)	34.8 (34.0;35.6)
Mean adjusted power	53.7 (52.7;54.7)	53.0 (52.0;54.0)	52.6 (51.6;53.6)	53.0 (52.0;53.9)
% of significant global treatment effect	38.6	38.5	38.3	38.8
% of significant treatment effect (ARDS)	25.7	24.6	25.5	25.1
% of significant treatment effect (No ARDS)	27.6	26.9	27.2	26.4
% of significant closed testings	12.0	11.0	12.2	11.2

Table 15: Ability of randomization procedures to show treatment's efficacy, 5000 uncorrelated datasets of size 300.
Between parenthesis : 95% confidence interval calculated using the central limit theorem

B.4 Scenario 2: 820 patients (Uncorrelated datasets)

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
MSE	0.18	0.18	0.18	0.18
RMSE	0.03	0.03	0.03	0.03

Table 16: Mean Squared Error and Root Mean Squared Error calculated for each randomization algorithm on 5000 generated uncorrelated datasets of size 820.

	MSB		Blocks (4)		Minimization (2 classes)		Minimization (4 classes)	
	No ARDS	ARDS	No ARDS	ARDS	No ARDS	ARDS	No ARDS	ARDS
Centers	0.04	0.02	0.00	0.00	0.06	0.08	0.06	0.02
Immunosupression	0.00	0.00	3.62	3.98	0.00	0.00	0.00	0.00
Age	0.00 (0.02)	0.00 (0.00)	4.54 (4.74)	4.98 (4.98)	0.02 (0.46)	0.12 (0.50)	0.00 (0.12)	0.00 (0.00)
Vasopressor dose	0.16 (0.00)	0.16 (0.00)	4.96 (4.42)	5.14 (5.38)	0.08 (0.76)	0.10 (0.74)	0.00 (0.08)	0.00 (0.08)
PaO2/FIO2	0.02 (0.00)	0.02 (0.00)	4.82 (4.66)	4.94 (4.50)	0.00 (0.12)	0.06 (0.12)	0.02 (0.02)	0.00 (0.02)
Serum lactate	0.80 (0.00)	0.62 (0.00)	5.04 (5.04)	4.78 (4.86)	0.26 (1.26)	0.06 (1.62)	0.00 (0.48)	0.00 (0.36)

Table 17: Percentage of p-value of stratified imbalance tests below 5% on the 5000 generated uncorrelated datasets of size 300. The imbalance tests are Wilcoxon signed-rank test and Student's t-test (between parenthesis) for quantitative covariates and chi-squared test for qualitative covariates.

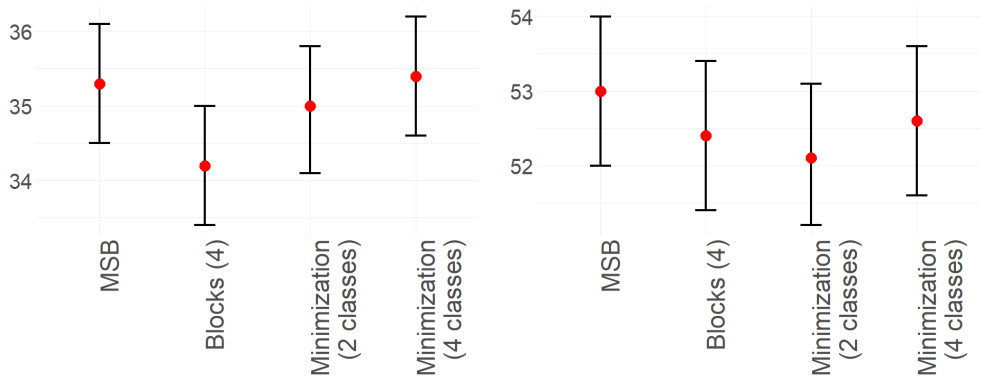
	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
ARDS	0.40	0.00	0.00	0.00
Centers	0.04	0.00	0.02	0.00
Immunosuppression	0.00	4.22	0.00	0.00
Age	0.02 (0.00)	5.18 (4.92)	0.36 (0.80)	0.00 (0.04)
Vasopressor dose	0.18 (0.00)	5.28 (4.58)	0.06 (0.80)	0.00 (0.10)
PaO2/FIO2	0.04 (0.00)	4.64 (4.74)	0.02 (0.16)	0.00 (0.00)
Serum lactate	0.68 (0.00)	4.78 (4.72)	0.02 (1.74)	0.00 (0.40)

Table 18: Percentage of p-value of global imbalance tests below 5% on the 5000 generated uncorrelated datasets of size 300. The imbalance tests are Wilcoxon signed-rank test and Student's t-test (in parenthesis) for quantitative covariates and chi-squared test for qualitative covariates.

	MSB	Blocks (4)	Minimization (2 classes)	Minimization (4 classes)
Mean unadjusted power	50.7 (49.8;51.7)	50.7 (49.7;51.6)	50.1 (49.2;51.1)	51.4 (50.4;52.3)
Mean adjusted power	66.2 (65.2;67.2)	66.4 (65.4;67.4)	66.0 (65.0;67.0)	66.7 (65.7;67.7)
% of significant global treatment effect	57.3	57.4	57.4	57.8
% of significant treatment effect (ARDS)	43.5	42.6	41.9	43.7
% of significant treatment effect (No ARDS)	44.8	45.7	45.0	46.0
% of significant closed testings	31.0	30.2	29.9	31.7

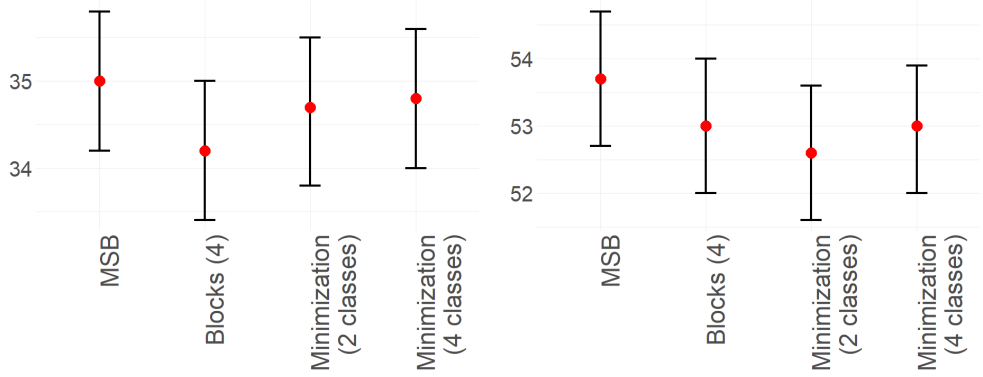
Table 19: Ability of randomization procedures to show treatment's efficacy, 5000 uncorrelated datasets of size 820.
Between parenthesis : 95% confidence interval calculated using the central limit theorem

B.5 Visual comparisons of adjusted and unadjusted power



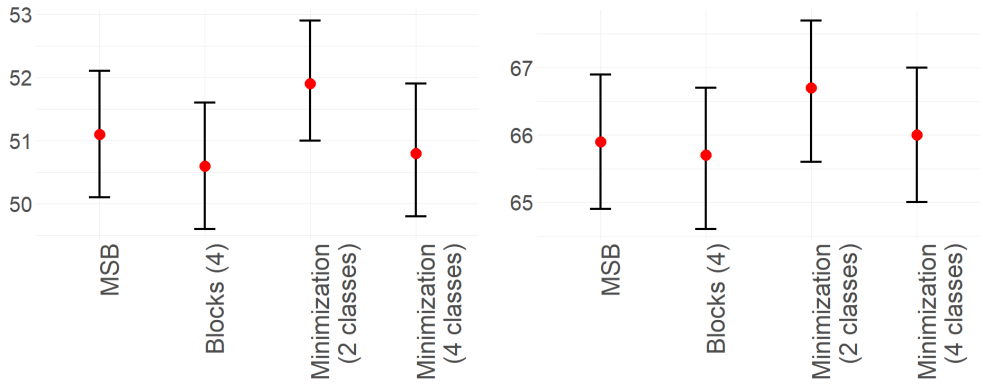
(a) Mean unadjusted power (b) Mean adjusted power

Figure 4: 95% confidence intervals obtained using central limit theorem for mean adjusted and unadjusted power, 5000 correlated datasets of size 300



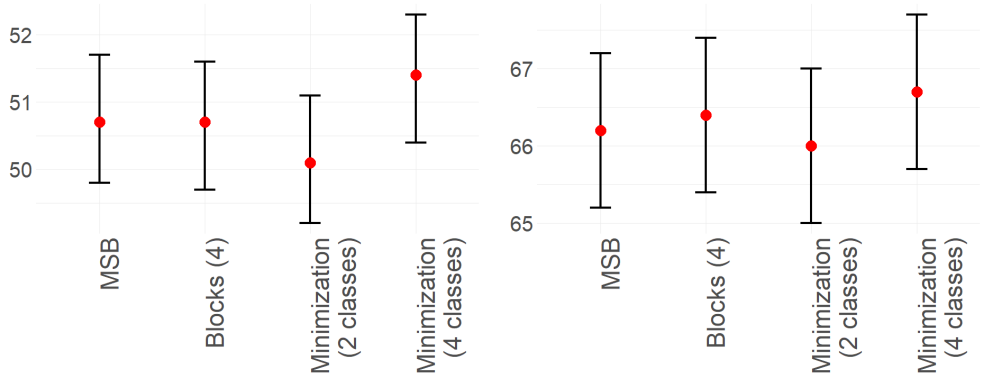
(a) Mean unadjusted power (b) Mean adjusted power

Figure 5: 95% confidence intervals obtained using central limit theorem for mean adjusted and unadjusted power, 5000 uncorrelated datasets of size 300



(a) Mean unadjusted power (b) Mean adjusted power

Figure 6: 95% confidence intervals obtained using central limit theorem for mean adjusted and unadjusted power, 5000 correlated datasets of size 820



(a) Mean unadjusted power (b) Mean adjusted power

Figure 7: 95% confidence intervals obtained using central limit theorem for mean adjusted and unadjusted power, 5000 uncorrelated datasets of size 820

B.6 Results of the factor analysis of mixed data

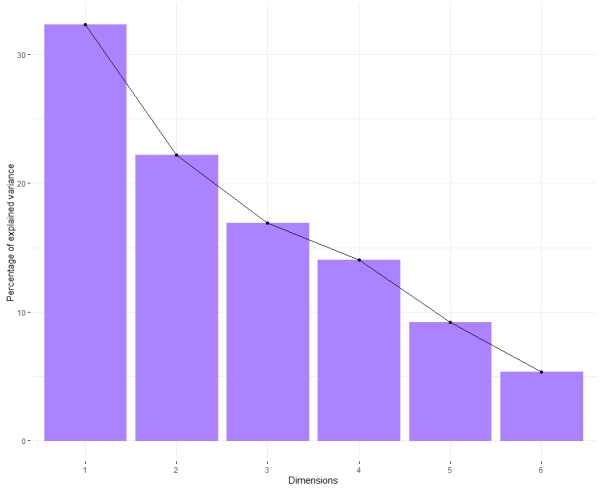


Figure 8: Barplot of the variance, Mixed Data Factorial Analysis conducted on the original dataset, missing values imputed with missMDA

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Age	14.47	1.25	0.10	82.79	0.00	1.40
PaO2/FIO2	33.41	6.59	1.94	9.05	4.84	44.17
Vasopressor dose	0.26	57.01	0.01	0.67	41.34	0.70
Serum lactate	0.29	3.86	89.35	0.11	6.38	0.00

Table 20: Contributions to the factors of continuous covariates, Mixed Data Factorial Analysis on the original dataset, missing values imputed with missMDA

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
ARDS	20.43	0.15	0.38	1.41	1.75	26.85
No ARDS	19.66	0.14	0.36	1.36	1.68	25.84
Immunosuppression	3.20	8.64	2.19	1.28	12.27	0.29
No immunosuppression	8.28	22.35	5.67	3.32	31.74	0.74

Table 21: Contributions to the factors of discrete covariates, Mixed Data Factorial Analysis conducted on the original dataset, missing values imputed with missMDA

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
Center 1	0.12	-0.18	1.24	0.19	-0.48	0.71
Center 3	-0.59	0.03	0.61	0.61	-0.09	1.82
Center 4	-1.14	-0.18	-0.26	-1.90	-0.94	-0.73
Center 5	-0.39	0.70	-0.15	0.04	0.54	0.68
Center 7	1.13	-1.00	-0.33	-0.28	0.34	-1.90
Center 8	-1.12	-1.36	-1.14	1.13	-0.15	-0.54
Center 9	0.90	0.93	-0.47	-0.09	0.04	-0.48

Table 22: V-test by center and dimension, from a Mixed Data Factorial Analysis with center as a supplementary variable