Distributive trials: a novel design to screen or evaluate multiple simultaneous interventions in clinical trials

Skerdi Haviari (✉ skerdi.haviari@aphp.fr)


France Mentré


Research Article

Keywords: Trial design, methodology, simulation

Posted Date: July 20th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3168385/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background: In some medical indications, numerous interventions have a weak presumption of efficacy, but a good track record or presumption of safety. This makes it feasible to evaluate them simultaneously. Here we introduce a modified factorial trial design that randomly allocates a pre-specified number of interventions to each participant, and statistically tests each intervention. We compare it to factorial trials, parallel-arm trials and multiple head-to-head trials, and derive some good practices for its design and analysis. Methods: We simulated various scenarios involving 4 to 20 candidate interventions among which 2 to 8 could be simultaneously allocated. A binary outcome was assumed. One or two interventions were assumed effective, with various interactions (positive, negative, none). Efficient combinatorics algorithms were created. Sample sizes and power were obtained by simulations in which the statistical test was either a difference in proportions or a logistic regression Wald test with or without interaction terms for adjustment, with a Bonferroni multiplicity-adjusted alpha risk in both cases. All code is provided without the need for compiling. Results: Distributive trials reduce sample sizes 2- to 7-fold compared to parallel arm trials, and increase them 1- to 2-fold compared to factorial trials, mostly when fewer allocations than for the factorial design are possible. An unexpectedly effective intervention causes small decreases in power (<10%) if its effect is additive, but large decreases (possibly down to 0) if not, as for factorial designs. These large decreases are prevented by using interaction terms to adjust the analysis, but these additional estimands have a sample size cost and are better pre-specified. The issue can also be managed by adding a true control arm without any intervention. Conclusion: Distributive randomization is a viable design for mass parallel evaluation of interventions in constrained trial populations. It should be introduced first in clinical settings where many undercharacterized interventions are potentially available, such as disease prevention strategies, digital behavioral interventions, dietary supplements for chronic conditions, or emerging diseases. Pre-trial simulations are recommended, for which tools are provided.

Introduction

As modern medicine expands, medical management options are constantly growing; for example, more drugs are approved (1, 2) than withdrawn (3), and thus the total available number is increasing. The same is true for medical devices (4), supplements (5) and even more so for non-pharmacological, non-device interventions (that can hardly be withdrawn in any meaningful sense). Thus, if only by repurposing, available options will inevitably come to include weak ones, with little pre-clinical evidence base and no randomized trials. This is especially true in areas such as chronic disease, functional symptoms, lifestyle interventions, and/or prevention strategies (e.g. different diet/supplement/exercise/meditation/education types). This can also be true when a new disease or disease form emerges (e.g. long COVID). If many possible interventions exist, and most of them are innocuous or easily manageable, it becomes feasible to administer them simultaneously, including to evaluate them in a trial, in order to quickly generate strong evidence.

One way to do this is using factorial trials: if there are K interventions, and each of them can be administered independently, $2^K$ combinations are defined and can each be tested. While extensively used for industrial or research experiments, this approach is rarely used in clinical trials beyond a 2x2 factorial
design (with 4 arms) (6). One of the main reasons is that it becomes complicated, both pragmatically and ethically, to administer many interventions to some patients and none to others. In this study, examining cases where $K$ is relatively large (4 and above), we introduce a limit $k$ to the number of interventions that can be administered simultaneously in each patient, and explore the most natural way to put it in practice, which is to administer $k$ interventions to each patient.

The basis for the limit $k$ can vary across settings: it can be physiological (fear of combining too many drugs for a new pathogen), ergonomic (avoid introducing too many changes to a digital interface or exercise routine), or even economic (in cases where participants incur costs for each intervention). Using the $\binom{K}{k}$ notation for the binomial coefficient, this creates $\binom{K}{k}$ different arms with $k$ interventions each. We call this distributive randomization, and compare distributive trials based on it to full factorial trials, "capped" factorial trials where combinations with more than $k$ interventions are excluded, parallel-arm trials, wholly separate trials, and a version of distributive trials with an additional true control arm without any intervention (Fig. 1).

In addition to the method of randomization, the analysis of the results is not obvious. In a full or a capped factorial trial, each intervention is evaluated in all treated versus all non-treated patients with this intervention, and an adjustment is made for multiplicity (testing several hypotheses at the same time), either explicitly or by using a test that takes it into account (such as ANOVA or Tukey's test). It can also be necessary to apply the same adjustment for multiplicity with parallel arms, which is often done with Dunnett's test, and even theoretically for separate trials, although for the latter there is no real way to do so across studies. We apply the same logic for distributive randomization as for standard factorial trials, and adjust for multiplicity. In cases where multiple efficacious interventions and/or interactions are expected, linear or logistic regression is the most natural choice (7), allowing the analysis strategy to be pre-specified to focus on the most relevant hypotheses.

Materials and methods

Combinatorics for allocation table computation

Because of the high number of treatment combinations and the need to run simulations for intricate cases, we needed efficient combinatorics algorithms to reduce the set of explicitly computed combinatorial allocations, and obtain an allocation table for simulations efficiently.

To this end, distributive randomization was computed as a subset of a table corresponding to a full factorial design, where, among $2^K$ combinations allocated with probabilities

$$\prod_{i=1}^{K} [X_i p_i + (1 - X_i) (1 - p_i)]$$
where \( X_i \) indicates the allocation of intervention \( i \), and \( p_i \) is the intervention's allocation probability, \( p_i = 0.5 \) for balanced allocations and \( p_i \neq 0.5 \) for unbalanced allocations), only \( \binom{K}{k} \) combinatorial allocations are allowed, and probabilities are normalized to sum to 1. This is equivalent to a situation where patients with more or less than \( k \) interventions would be rerandomized until they received \( k \) interventions.

To reduce the number of explicit combinations, the following reasoning was used: “interesting” interventions were defined as either non-null efficacy and/or unbalanced allocations. First, for the subset of \( L \) interesting interventions, the full table with all combinations was computed, with the corresponding probability for each of the \( 2^L \) combinations (as a partial draw) explicitly computed. Formally, if we number interventions so that 1 to \( L \) denotes interesting ones and \( L+1 \) to \( K \) denotes uninteresting ones, we have for the subset of interesting interventions

\[
p'(X_I) = \prod_{i=1}^{L} (X_i p_i + (1 - X_i) (1 - p_i))
\]

where \( X_i = \{X_1, X_2, X_3, ..., X_L\} \) is the vector of indicator variables for the allocations of interesting interventions, \( L \) is the number of interesting interventions and \( p_i \) is as above.

Then, the draws being independent, the probability of each of the \( 2^L \) combinations of interesting interventions in the final allocation is obtained, by multiplying the probability of that combination as a partial draw, \( p'(X_i) \), by the probability of getting exactly the correct number of allocations among the remaining uninteresting interventions (by the binomial distribution with \( p = q = 0.5 \) by definition of uninteresting interventions), \( p''(X_I) \), then normalizing to obtain the final probability of allocation, \( p(X_I) \).

\[
p'(X_i) = \sum_{i=1}^{L} X_i
\]

\[
p''(X_I) = \binom{K-L}{k-1} 0.5^{K-L}
\]

\[
p'''(X_I) = p'(X_I) \times p''(X_I)
\]

\[
p(X_I) = \frac{p'''(X_I)}{\sum p'''(X_I)}
\]

The convention \( \binom{Y}{y} = 0 \) for \( y < 0 \) or \( y > Y \) is used when computing \( p''(X_I) \) to give arms with too many or too few allocations a probability of 0. In the provided code, to enable others to explore cases where ineffective allocations have arbitrary allocation ratios (and thus are not equiprobable), all possible values of \( p''(X_I) \) are pre-computed by explicitly drawing the \((L+1)\)th to \( K\)th interventions one by one and keeping track of the probabilities for each overall number of allocations.

For the capped factorial the computations were as above except \( p''(X_I) = \sum_{a=0}^{k} \binom{K-L}{a-1} 0.5^{K-L} \) since fewer than \( k \) allocations are also allowed.
For the full factorial trial computations were as above except $p''(X_i)$ was set to 1 since any number of uninteresting interventions is allowed.

For the controlled distributive design, an arm with no interventions was added at the end of the computation, with a specified probability, and the other $p(X_i)$ terms were adjusted.

For parallel-arms each arm was computed explicitly and the allocation ratio was 1:1 between each intervention and control.

**Adding outcomes to the allocation table and simulating a trial**

The probability of clinical success was computed for each of the $2^L$ combinations of interesting interventions (ignoring the uninteresting ones, which have no effect by definition). This was done by applying an explicit list of successful outcome probabilities, with a priority for specified higher-order combination (e.g. if intervention 1 + intervention 2 combined had 100% success, and intervention 1 had 70% success, the former value was used for all arms containing 1 and 2, and the latter for all other arms containing 1). The list used in the scripts was written to avoid any ambiguous or implicit cases.

The full allocation table with outcomes allowed overall success probabilities by presence or absence of interesting interventions to be readily computed, and enabled fast simulations by drawing patients and their outcomes in the detailed combinatorial arms. To complete the design matrix with the uninteresting interventions, for each simulated patient the remaining draws were made:

- for distributive randomization, $k-l$ uninteresting interventions among $K-L$ were drawn at random
- for the full factorial design, interventions among the remaining $K-L$ were drawn independently
- for the capped factorial design, in this study no simulations were run, so no algorithm was devised to complete the design matrix (in this case, it would need to keep track of the probability for each number of uninteresting interventions, and first draw that number, then their identity)

**Statistical risk, sample size and power computations**

Type I ($\alpha$) risk was set at bilateral 5% and type II ($\beta$) at 10%. For consistency, due to its generic nature and its applicability to all statistical tests, a Bonferroni adjustment was applied for multiplicity in all situations where the effect of several interventions was tested at the same time, regardless of the trial design. The correction was applied to the $\alpha$ threshold, which was divided by the number of tests. The corrected $\alpha$ risk was then used for sample size and power computations. The $\beta$ risk was not adjusted : e.g. if several interventions were equally effective, $\beta$ was the probability of missing the first one, regardless of the others.

In the base case, only one intervention was considered effective, and therefore analytical formulas were used for sample size and power, based on difference of proportions under a Gaussian approximation (8–11), using the previously obtained table. The formula is
where \( r \) is the ratio between subjects not receiving and receiving the tested intervention (given by the trial design), \( n_I \) is the number of subjects receiving the tested intervention, \( r \) the number not receiving it, \( n \) is the total sample size, \( p_I \) is the success probability for those receiving the intervention and \( p_C \) for those not receiving it, and \( z_\alpha \) and \( z_\beta \) denote the standard normal values for the desired type I and type II errors.

In more complicated cases, where multiple interventions could be effective, with or without interactions, analysis was performed by logistic regression and simulations were used for trial size and power. Power computation was straightforward; using the previously derived table to draw allocations and outcomes for virtual patients, 5000 trials were simulated, and power was approximated as the frequency of a \( p \)-value below the \( \alpha \) threshold.

Trial sizes by simulation were obtained with a bespoke algorithm. The probability of discovery (rejection of the null) was modeled using a modified probit model with a single variable, the square root of the trial size. This specification was used because many closed-form sample size formulas are of the form

\[
N = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{K}
\]

where \( K \) is an expression of the effect size; expressing power as a function of sample size thus gives an equation of the form \( p_d = \Phi(a + b\sqrt{N}) \) with \( a \) and \( b \) as parameters to optimize, and \( p_d \) the probability of a discovery (rejection of the null).

The modification of the probit model was the use of the following reparameterization:

\[
p_d = \Phi \left(z_{1-\beta} + b(X - X_0)\right)
\]

where \( X \) is the square root of the trial size, \( b \) is a nuisance parameter, \( X_0 \) is the parameter of interest (the square root of the right trial size), and \( z_{1-\beta} \) ensures that when \( X \) is equal to \( X_0 \), the desired power is obtained.

Sampling of trial sizes was started with one (assumed) non-significant trial with 1 patient, one (assumed) significant trial with \( 10^6 \) patients, and 48 trials between 50 and 10000 patients (with equal log-scale increments). The 48 initial trials were re-run if they were either all significant or all non-significant, then additional trials were simulated in batches of either 50 trials or 10\% of already simulated trials (whichever highest), with new fits of the probit model after each batch, until at least 5000 simulations were run.

Simulated sample sizes for each batch were spread uniformly in an interval of \( \hat{X}_0 \) (current estimate) \( \pm \) its standard error derived from the Hessian matrix, which narrowed at each step, or \( \pm 10 \), whichever was smallest. This quickly resulted in sampling occurring only around the desired value, alleviating misspecification concerns.
As a validation step, computations of sample size for Fig. 3B (2 to 8 allocations among 3 to 20 interventions) were all checked with 5000 simulations each and showed an observed power of 90.02 ± 0.5857%, whereas 90 ± 0.5888% would be expected simply due to chance if starting from the known true value and adding sampling variances (first from the sample size algorithm, which usually stopped at 5400 simulations, then from the check with 5000). Outliers could very rarely occur (apparently due to wildly improbable first batches) and were re-run after being diagnosed visually (future implementations could benefit from a mechanism to “forget” the first batches).

**Assumptions and simulation scenarios**

In all studied scenarios, the rate of successful outcomes was 50% in the absence of any intervention, and most interventions were ineffective, having no impact on the success rate. When shown, factorial trials used a 1:1 randomization, violating the individual-patient limit of k interventions of distributive trials for some of their patients, and are presented as an ideal but unfeasible benchmark.

Table 1 shows an overview of all scenarios and their progression. Briefly, we start from a simple case with a single effective intervention that needs to be found in the trial, and show sample size gains with distributive trials. We then move to a situation where this single-effective assumption is false, changing power (if incorrectly foreseen) or sample size (if correctly foreseen) depending on interactions between the two effective interventions. Finally, we try different design and analysis modifications, requiring decreasing levels of *a priori* knowledge, to detect both interventions with as few patients as possible. In scenario group 1, only one intervention was assumed to possibly be effective, with successful outcomes going from the baseline 50–70% with that intervention (regardless of any other interventions in those patients). Sample sizes were obtained under closed-formed formulas for a difference of proportions with unbalanced arms shown above, pooling all patients who did not receive a particular intervention to serve as control for that intervention (trading power for a small bias against the other interventions in distributive designs). For parallel arm trials, a non-pooled analysis is also shown (each intervention versus the control arm), as it is the most intuitive and common analysis. For distributive and capped factorial designs, the number of simultaneously feasible interventions, denoted k, was either 2, 4 or 8, and the total number of candidate interventions in the trial, K, varied from k + 1 to 20.
<table>
<thead>
<tr>
<th>Scenario group</th>
<th>Designs</th>
<th>k</th>
<th>K</th>
<th>Interventions’ success probability</th>
<th>Analysis</th>
<th>Foresight†</th>
<th>Fig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Int. 1</td>
<td>Int. 2</td>
<td>1 + 2 combined</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>All</td>
<td>2</td>
<td>4</td>
<td>70%</td>
<td>50%</td>
<td>70%</td>
<td>Difference of proportions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>3–20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Distributive, factorial</td>
<td>3–8</td>
<td>5, 10, 15, 20</td>
<td>70%</td>
<td>50%</td>
<td>50%</td>
<td>Logistic regression without interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2, 4, 6, 8</td>
<td>3–20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2–8</td>
<td>3–20</td>
<td>70%</td>
<td></td>
<td>84.5%*</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60% 70%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Distributive, factorial</td>
<td>2–8</td>
<td>3–20</td>
<td>70%</td>
<td>70%</td>
<td>84.5%*</td>
<td>Logistic regression without interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60% 70%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Distributive, factorial</td>
<td>3–8</td>
<td>5, 10, 15, 20</td>
<td>70%</td>
<td>50%</td>
<td>50%</td>
<td>Logistic regression with pre-specified interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2, 4, 6, 8</td>
<td>3–20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2–8</td>
<td>3–20</td>
<td>70%</td>
<td></td>
<td>84.5%*</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60% 70%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>(controlled)</td>
<td>4</td>
<td>10</td>
<td>70%</td>
<td>70%</td>
<td>84.5%*</td>
<td>Difference</td>
</tr>
<tr>
<td>Distributive with 0–80% intervention-free arm</td>
<td>2</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70% of proportions or logistic regression with interaction either absent or backward selected or pre-specified</td>
<td>6B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>84.5%*</td>
<td>6C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70%</td>
<td>6D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 Simulation and analysis scenarios

50% success means no intervention effect (this is the baseline in all scenarios) * denotes logit-scale additivity, i.e., expit(2ln(7/3)), which is rounded to 84.5% in the table and 84.48276% in the code implementation † "foresight" denotes whether trial size was computed with true assumptions; if so, the quantity of interest is sample size, if not, the quantity of interest is power (i.e., how it is affected by those wrong assumptions).

In scenario group 2 the comparison was restricted to distributive versus full factorial designs. A first subset of this scenario group explored how k and K impacted sample size, with the same settings as above except for the following:

- the analysis relied on logistic regression (allowing detection of, and adjustment for, more than 1 effective intervention), with main effect terms only

\[
\text{logit} \left( p \left( Y_i = 1 \right) \right) = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_K X_{K,i}
\]

where \( Y_i \) is the indicator variable for success or failure for subject i, \( \alpha \) is the intercept, \( \beta_j \) is the coefficient for intervention j, and \( X_{ij} \) is the indicator variable for whether subject i received intervention j.

- the sample sizes were therefore obtained by simulation

- holding K fixed at 5, 10, 15 or 20, all possible values of k were tested to find the ideal k/K ratio for sample size minimization

- holding k fixed at 2, 4, 6 or 8, K was increased from k + 1 to 20 to examine the incremental cost of evaluating additional interventions (e.g. with weak presumption of efficacy) in terms of sample size

In a subset of scenario group 2, the simulation truth did not match the sample size assumption: a second intervention was effective, while only one was expected to be. The impact of this incorrect assumption on power was computed by simulation. The second intervention's efficacy, unplanned for at the sample size computation stage, was one of the following four:

- same main effect as the expected effective intervention (70% success when administered standalone, regardless of other interventions except the other intervention with a main effect), with additivity on the logit scale (the combination had an 84.48276% success rate)
- same main effect as the expected effective intervention, but strong synergy (100% success rate for the combination)

- same main effect as expected effective intervention, but no additivity of effects (the combination also has a 70% success rate)

- smaller main effect than the expected effective intervention (60% success rate rather than 70% previously), and not additivity of effects (the combination also has a 70% success rate)

In scenario group 3 similar interaction assumptions as for scenario group 2 were used, except they were taken into account initially to compute trial size.

The next scenario group explored a strategy to increase robustness with respect to interactions between effective interventions. While it is generally a bad idea to run factorial trials when interactions are expected (7, 12), and the argument should apply to distributive trials, wrong assumptions on interactions are possible. Thus, some robustness against them can be useful to limit unexpected loss of power, even at some upfront sample size cost. Therefore, in scenario group 4, the same assumptions as scenario group 2 were used, but an additional estimand, a pre-specified interaction term between the two effective interventions, was included in the logistic model.

$$\text{logit} \{p(Y_i = 1)\} = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_K X_{K,i} + \beta_{12} X_{1,i} X_{2,i}$$

with the same notations as above, except 1 and 2 are the two effective interventions and $\beta_{12}$ is the additional interaction term.

Finally, in scenario group 5, we explored additional strategies to reduce the sample size and stay robust against interactions. The addition of a true control arm representing 0 to 80% of patients was tested, as well as a set of different analysis algorithms:

- logistic regression without any interaction term
- logistic regression with a pre-specified interaction term between the two effective interventions
- logistic regression with a backward elimination strategy for interaction terms, starting from all interactions involving the analyzed intervention and removing those with $p > 0.05$
- same as above but with a threshold at $p > 0.25$
- pooled difference, testing the difference in success proportions in intervention versus non-intervention subjects (and accepting some bias to decrease sample size)

In this scenario group, we held other parameters at constant values (previously explored):

- $k$ and $K$ were either $\{k, K\} = \{4,10\}$ or $\{k, K\} = \{2,20\}$
success rate was 70% for two interventions and 50% (no effect) for others, and the former’s combined effectiveness was either 84.48276% (logit-scale additivity) or 70% (lack of additivity).

**Software**

All computations were made with R version 4.2.2.

**Results**

As shown in Fig. 2 (panels A, B, E and F), when analyzed by testing differences in proportions, distributive trials perform as well as factorial trials when k = K/2 and much better than parallel arm trials. The capped factorial design behaves similarly to the distributive one. The parallel arm design requires 1.5-2x more subjects than the distributive one, even in the least efficient form of the latter, when k = 2, and with the most efficient (and biased) analysis for the former, pooling arms in a large control. Increasing the number of per-patient interventions to 4 and 8 (Fig. 2, panels E and F) allowed further reductions in sample size, approaching that of the full factorial design.

Still assuming a single effective intervention, but switching the analysis to logistic regression to make it possible to find multiple effective interventions without bias, Fig. 3A shows that the most efficient number of allocations per patient is 50% of candidate interventions, bringing trial sizes close to those of factorial trials (background grey lines for 1:1 factorial design in all panels). Figure 3B shows sample sizes for a given number of allocations, depending on how many interventions are to be tested. The switch to logistic regression, while reasonable if several interventions can be effective, is costly in terms of sample size. The latter is approximately 80% higher compared to a pooled difference of means (Fig. 3B purple line versus Fig. 2C red line), but remains way below those of a parallel trial (from 2x to 5x depending on k and K for K < 20, and even more if K is larger). Incremental increases for additional candidate interventions to be tested are also small compared to an extra arm in a parallel trial, and smaller as k increases (+ 220 subjects for a 20-arm Bonferroni-adjusted parallel trial versus around + 120 with k = 2 and + 20 for k = 8). Sample size increases for additional candidates are very small in particular if the k/K ratio can be maintained (Fig. 3A, distance between curves for 5 extra candidates), much like a factorial trial.

The assumption on the number of effective interventions may be false, which will impact power; Fig. 3C-F shows the drop in power when only one intervention was expected to be effective, but another one was also (for a total of two effective interventions). The extent of the decrease depends heavily on the nature of the interaction between the two effective interventions, from small (< 15%) for a logit-scale additive effect (Fig. 3C) to non-existent and even reversed for a synergistic effect (Fig. 3D) to large for a lack of additivity (i.e., a negative interaction) (Fig. 3E), with the effect being somewhat mitigated if there is still some contrast between the interventions and k is small (Fig. 3F). The decrease is modest compared to that observed with a factorial design (background gray lines) if the effective interventions are sufficiently diluted among ineffective ones. Note that power can be 0 in the strong synergy case (Fig. 3D), due to infinite standard errors when all interventions except one are allocated to each patient, and the combination has a perfect success rate (an unmistakable finding that would presumably lead to an adaptation of the analysis in a
real-world setting). The same thing happens when all interventions but one are allocated to each patient and there is no additivity (Fig. 3E), because in that case all patients have the increased success rate (70%).

When multiple effective interventions are expected, sample size can be adapted to maintain statistical power. Figure 4 shows that increases in sample size are modest unless a large fraction of interventions is given to each patient, the effect is non-additive, and the main effects have exactly the same size (Fig. 4C). This special case is difficult, even for factorial trials (Fig. 4C, background gray line).

While positive interactions increase power and do not truly create problems, negative interactions decrease power. Figure 5 shows how adding an interaction term in the logistic model can allow it to adjust for the negative interaction and prevent loss of power with an extra effective intervention. It shows the same cases as Fig. 3, but with an added interaction term used only for adjustment (conclusions are only drawn about main effects).

The best ratio \(k/K\) is decreased to 40%, and the incremental cost of an additional candidate intervention remains small when \(k/K\) is maintained (Fig. 5A). The overall sample sizes are very similar to those without an interaction term (Fig. 5B). Note the irregularities in sample size when all but a few interventions are allocated (Fig. 5A and 5B), which hinders identifiability. This scenario, however, likely makes distributive randomization useless, because if one can allocate almost all interventions to each patient, a factorial trial will almost always be a better choice. Power is generally maintained regardless of the nature of the interaction, except when all interventions but one are allocated (Fig. 5C-F), and when there is an unexpected intervention with synergistic effects and a (too) high \(k/K\) ratio (Fig. 5D). Sample size shifts when making the correct assumption, as in Fig. 4, were small (as follows from power not varying) and are not shown.

The previous analysis method only adds one interaction term to adjust for a suspected negative interaction, but does not provide guidance on how to plan for completely unexpected negative interactions. Some possible mitigation strategies are shown in Fig. 6. Figure 6A and Fig. 6B show different design and analysis strategies for 4 allocations among 10 candidates, of which 2 are effective either with logit-scale additivity (i.e., no interaction) in Fig. 6A, as in Fig. 5C, or without additivity (i.e., negative interaction) in Fig. 6B, as in Fig. 5E. The allocation of some patients to a true control arm without any active intervention, which corresponds to the controlled distributive design from Fig. 1, can lead to a substantial sample size decrease.

As previously, a pre-specified interaction term is helpful to adjust for a negative interaction, but requires a good intuition before the trial. Trying to circumvent the need for good intuition by using a backward elimination strategy, starting from all interactions involving an intervention and iteratively eliminating those with \(p > 0.25\), also has limited utility, but a more stringent threshold (\(p > 0.05\)) can work, at some sample size cost. Adjusting on all interactions involving the analyzed intervention is not viable because it requires many more subjects (> 9000, not shown on the graph). The effect is even worse for all two-way interactions and not only those for the analyzed intervention (> 30000, also not shown). An iterative backward elimination strategy starting from all interactions, is able to prune them and bring sample sizes to reasonable levels, very similar to starting from interactions only with the analyzed intervention (not shown). Figure 6C and 6D
show sample size determinations for the same true effects but with a strategy involving 2 allocations among 20 candidate interventions; here, interaction terms are not very useful because effective interventions are “diluted” among non-effective ones.

The red lines in Fig. 6 show a strategy where interventions are analyzed by difference of means while simply pooling receiving versus non-receiving subjects, as in Fig. 2, and accepting that the difference will be biased by the mean of the other interventions; the strategy performs surprisingly well in this setting with few effective interventions (2 effective among either 10 or 20), bested only by a combination of a small control arm and a pre-specified interaction term in case of no additivity (Fig. 6B), but its performance would likely quickly drop if many candidate interventions have similar efficacy, and may also run into problems with unbalanced designs (7).

**Discussion**

This work introduces a new trial design to evaluate many interventions in parallel, deliberately confounding the effects in the same subjects while adjusting the analysis to retrieve correct results. The philosophy is similar to that of a factorial trial (the possible combinations per subject are a subset thereof) but mitigates the practical issues with the high number of per-patient allocations in factorial trials beyond the 2x2 case, while still retaining large decreases in sample size compared to parallel trials. It can also have the added benefit of providing (multiple) active interventions to all subjects, which could increase participation rates by as much as 20–50 percentage points in clinical trials given patients’ distaste for control arms (13, 14).

Many biomedical fields could benefit from distributive randomization, and there are several examples of recent trials that could have used it. During the early COVID pandemic, several in vitro screening experiments produced many candidates for drug repositioning, many of which did not appear to be effective with further research (15). Trials such as DisCoVeRy (16), RECOVERY (17) or REMAP-CAP (18) could have produced faster results or evaluated more candidates with such an approach. Our experience with DisCoVeRy also showed high rates of participant refusal due to the control arm. In another setting, the NUDGE-FLU trial recently evaluated 9 different electronic letters to encourage influenza vaccine uptake in the general population (19), using a parallel-arm design, and could have similarly benefitted from multiple interventions per subject; only 2 out of 9 arms showed a (small) effect, and the planned instrumental variable analysis could have benefitted from more interventions given to each subject (and so would the subjects themselves). Future similar prevention trials could benefit from distributive randomization. Nutrition is another field where many possible interventions can be tested and where the scattershot approach of traditional trials has difficulty producing results, due to unclear rationales and probably small effect sizes (20); large distributive randomization trials testing 20 or 30 nutrients in a few thousand patients could be of great help there, especially if linked with administrative databases for simpler follow-up. Finally, another area of interest is open online trials, which are starting to be a modality (21) and for which sample sizes are constrained and subject motivation is critical, both problems being alleviated by this design.
When planning a distributive trial, assumptions must be carefully laid out. In particular, one should avoid excessive optimism about the effectiveness of candidate interventions, as the history of clinical trials warrants. Even thoroughly vetted drug candidates in phase III fail due to a lack of efficacy approximately half the time (22). This is even more common with supplements (23, 24). Behavioral interventions, even without a clinical endpoint, also have a high failure rate (25). Unless a true control arm is added, distributive randomization actively relies on some interventions not working, and we fear that this could be a barrier to adoption; one way to manage this objection is simply to evaluate additional candidate interventions that are unlikely to work, because the incremental sample size increases are small.

There are many moving parts in the design of a distributive trial, some of which are introduced here (the k/K ratio, the existence and size of the control arm, the inclusion of interaction terms), and some of which remain to be addressed in the future or for particular cases (unbalanced allocations, forbidden or enriched combinations, dose levels). Because of these, one should run careful simulations when planning such a trial, and try to make sure the design is robust if the interventions do not work as expected. As a starting strategy, we recommend avoiding interventions with suspected non-additive effects due to e.g. the same mechanism of action (or, if impractical, adjusting with a few dedicated pre-specified interaction terms, at the cost of additional inclusions, or making the allocations mutually exclusive), increasing allocations per patient if possible (no further than 40–50%, less if negative interactions are suspected and cannot be adjusted), and admitting potentially ineffective candidate interventions in the trial liberally. Even outright ignoring interactions and adjustments becomes possible if effective interventions are diluted among more exploratory ones. A small true control arm (5–20% of subjects), while it introduces inequality between subjects and may be hard to justify to them and to investigators, can reduce sample size significantly; its presence should be debated case by case, as a tradeoff between cost and length on one hand, and fairness and absence of self-selection of subjects on the other. All scripts used for this study are provided, and the authors can be contacted to replicate or adapt the computations for a planned trial.

Conclusion

Distributive trials are an efficient new design that can help expedite evaluation in under-evidenced areas of medicine, such as emerging diseases, supplements, nutrition, probiotics, behavioral or lifestyle modifications, and digital interventions. Its strengths include more fairness to participants, small sample sizes, and low sensitivity to additional interventions with unlikely efficacy, while its adjustable parameters require foresight and simulation work from investigators. The aforementioned strengths are especially well-suited to a rapidly evolving healthcare landscape where therapeutic options are ever more numerous, digital tools are gaining in strength and reach and enabling larger sample sizes, patients are increasingly willing to challenge expertise and try things on their own, and the distinction between wellness and medicine can become blurry.

In this changing context, the clinical research community should ensure that the evidence-based spirit of modern medicine stays in step with patients and society and not confined to academic and industrial settings. Instead of often simply shrugging and citing lack of evidence, we can use innovative and efficient trial designs such as this one to produce it. Creativity in this area can help unlock the potential of not only
new inventions, but also nonpatentable or otherwise overlooked strategies that may have been available for years or even centuries, and about which evidence-based medicine’s lack of real answers is often not accepted by segments of broader society.

**Declarations**

**Supplementary materials**

Supplementary materials include all R scripts used to generate and display the presented data. Additional methods are available where parts of the code go further than what is presented in the main article.

**Ethics approval and consent to participate**

Not applicable as the study is based on mathematics and computer simulations.

**Consent for publication**

Not applicable as the study is based on mathematics and computer simulations.

**Availability of data and materials**

The exact code used to generate the simulations and the figures is provided in the Supplementary Appendix.

**Competing interests**

The authors declare that they have no competing interests.

**Funding**

No specific research grant was sought or obtained for this study. The authors would like to thank Université Paris Cité and APHP for providing the necessary academic freedom to complete it.

**Authors' contributions**

S.H. proposed the initial idea, wrote the computer code and drafted the manuscript. F.M. reviewed the biostatistical methodology and adapted the scenarios. Both authors reviewed the manuscript.

**Authors' information**


**References**


Figures
Trial strategies

Distributive trials versus other possible designs for parallel evaluation

Out of a pool of $K$ candidate interventions (here $K=4$)...

... $k$ maximum concomitant allocations are considered reasonable (here $k = 2$)

Figure 1

Distributive trials versus other possible designs for parallel evaluation
Figure 2

**Trial sizes for a single effective intervention** among 4 candidates (panel A) or among 3 to 20 candidates (panels C and D, same colors as in A) with either 2 (panels A-D), 4 (E) or 8 (F) allocations per patient. Panel B shows the total trial size ratio compared to the parallel-arm design (each intervention analyzed versus control) and the associated gains, with the same color scheme as in A. Panels E and F show improvements of the distributive and capped factorial design as more interventions (4 and 8 respectively) are tested per patient (same colors as in A). The efficacy assumption is that a single intervention brings the probability of...
a good outcome from 0.5 to 0.7, and the other interventions have no effect. Alpha risks (nominal 5%) are
Bonferroni-adjusted for multiplicity, and 90% power is sought. *the full factorial design is shown for
reference but ignores maximum allocations

**Figure 3**

Sample sizes using logistic regression and loss of power with an extra unexpected effective intervention
Panels A and B show the expected sample size with a single expected effective intervention that increases
clinical success from 50% to 70%, based on 5000+ simulations, using logistic regression for analysis and aiming for 90% power. Then, each panel shows a different scenario with an additional unexpected effective intervention with the same main effect size, and, for the combination, either: logit-scale additivity (the combination yielding 84.48276% clinical success) (panel C), strong synergy with 100% clinical success (panel D), or no additivity at all (panel E). In the final case, the unexpectedly effective intervention only has a 60% success rate, and no additivity (panel F). For each panel, the background gray line shows the sample size or true power of a factorial trial powered for the same situation with the same wrong assumptions (changes are mostly due to the multiplicity adjustment). For the distributive designs, each curve is for a different number of allocations per patient k (equal to its starting point minus 1), with one color hue per number of allocations.
Figure 4

**Increase in sample size with an extra expected effective intervention** Each panel shows the same scenarios as in Figure 3, but this time the scenario was correctly assumed to compute sample size.
Figure 5

Adjustment with an interaction term. Panels are the same as in Figure 3; the only difference is that an interaction term between the two effective interventions is fitted when performing the logistic regression.
Figure 6

**True control arms and mitigation strategies for interactions** Panel A shows sample sizes for a scenario with 2 effective interventions (70% success) with additive efficacy on the logit scale (84.48276% clinical success), 8 ineffective interventions (no change), and 4 allocations per patient, as a function of the size of the control arm (relative to total inclusions). Panel B shows the same situation but with non-additive efficacy (70% success for the combination, same as each standalone intervention). Panels C and D show the same two effective interventions (with and without additivity) but with 2 allocations among 20
candidates rather than 4 among 10. The analysis algorithms are: “Confound” for a pooled difference of means between treated and non-treated, accepting bias due to pooling but counting on a small proportion of effective interventions, “No interaction” for a simple logistic regression with the main effects, “Pre-specified” for a single interaction term between the two effective interventions, “Backward” for a backward elimination strategy starting from all interaction terms involving the analyzed intervention and keeping only those with a p-value below the threshold.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- Appendixv10BMCMRM.docx