mathematics

MDPI

Article

Optimal Weighted Multiple-Testing Procedure for Clinical Trials

Hanan Hammouri *, Marwan Alquran, Ruwa Abdel Muhsen and Jaser Altahat

Department of Mathematics and Statistics, Jordan University of Science and Technology, Irbid 22110, Jordan; marwan04@just.edu.jo (M.A.); rmabedalmohssen17@sci.just.edu.jo (R.A.M.); jyaltahat14@sci.just.edu.jo (J.A.)
* Correspondence: hmhammouri@just.edu.jo

Abstract: This paper describes a new method for testing randomized clinical trials with binary outcomes, which combines the O'Brien and Fleming (1979) multiple-testing procedure with optimal allocations and unequal weighted samples simultaneously. The O'Brien and Fleming method of group sequential testing is a simple and effective method with the same Type I error and power as a fixed one-stage chi-square test, with the option to terminate early if one treatment is clearly superior to another. This study modified the O'Brien and Fleming procedure, resulting in a more flexible new procedure, where the optimal allocation assists in allocating more subjects to the winning treatment without compromising the integrity of the study, while unequal weighting allows for different samples to be chosen for different stages of a trial. The new optimal weighted multiple-testing procedure (OWMP), based on simulation studies, is relatively robust to the added features because it showed a high preference for decreasing the Type I error and maintaining the power. In addition, the procedure was illustrated using simulated and real-life examples. The outcomes of the current study suggest that the new procedure is as effective as the original. However, it is more flexible.

Keywords: statistical algorithm; sequential group test; O'Brien and Fleming; Type I error and power; simulations; SAS software; hypotheses testing; biostatistics; categorical data analysis

MSC: 92B15; 62L10; 62L12; 62L15; 62P10

Citation: Hammouri, H.; Alquran, M.; Abdel Muhsen, R.; Altahat, J. Optimal Weighted Multiple-Testing Procedure for Clinical Trials. *Mathematics* **2022**, *10*, 1996. https://doi.org/10.3390/math10121996

Academic Editor: José Antonio Roldán-Nofuentes

Received: 15 May 2022
Accepted: 7 June 2022
Published: 9 June 2022

1. Introduction

Experimental designs were introduced firstly in agricultural projects. Industrial and laboratory researchers were inspired to adopt them into clinical trials of pharmaceuticals in humans because of their distinction in monitoring the experimental process to minimize errors. A clinical trial's primary purpose is to find the optimum medical treatment by comparing the benefits of competing therapies at minimal costs and within a short period. Doing so with the least possible errors is highly critical [1].

The first clinical trial designs utilized classical experimental designs. According to medical and physiological science progress, there was a need to change some elements during the trial process. The sample size could be modified, some trials could be terminated early, or the trial stages adjusted; the classical analysis cannot accommodate these modifications. Therefore, adaptive designs have been developed. An adaptive design is a method that involves modification of a current trial's design or statistical procedure in response to the data generated from the trials. In addition, they allow investigators to identify the best treatment under study without compromising its validity and integrity [2]. The types of adaptive design methods generally considered in clinical trials include an adaptive treatment switching design, a group sequential design, a biomarker-adaptive design, an adaptive randomization design, a drop-the-losers design, a sample size re-estimation design, and a hypothesis-adaptive design [2].

An interim analysis (a group sequential design) is one of the most popular options. Using interim analysis has several benefits that can be loosely categorized into ethical,

administrative, and economic categories. On the other hand, a group sequential design is morally imperative to control the results of clinical trials, including those involving human subjects, to minimize individual risks [3].

Since group sequential testing procedures became widely used in medical experiments. The procedures were crucial to be developed by researchers to generalize and modify procedures for varying circumstances. By way of example, Maurer and Bretz developed a class of group-sequential-weighted Bonferroni procedures with multiple endpoints, for which the correlations of sequential statistics are used. Consequently, the power was increased, while the family-wise error rate was effectively controlled [4]. On the other hand, Yiyong Fu presented follow-up work to Maurer and Bretz, proposing a Holm-type step-down exact parametric procedure for situations in which correlations are unknown. Further, he briefly outlined an extension of the partially parametric Seneta–Chen method that is naturally a group sequential design [5]. Zhenming Shun examined an approach that combined sample size re-estimation, a negative stop (stochastic curtailment), and group sequential analysis in a single interim analysis conducted with normal data [6].

Urach and Posch also considered an approach to improve critical boundaries for multi-arm experiments by using multi-arm group sequential designs with a simultaneous stopping rule. The resulting designs are also intended to optimize their boundaries' shape and determine their operating characteristics [7].

Most researchers recently adopted well-known procedures, such as the O'Brien and Fleming, Pocock, and Haybittle–Peto procedures [8].

For the development of group sequential testing procedures, O'Brien and Fleming (1979) provided a primary inspiration in this field [9], since they presented a direct and valuable group sequential testing procedure to compare two treatments in clinical trials. When one treatment performs better than the other, the trial is terminated using a smaller sample, where the procedure offers the same Type I error and power rates as a fixed one-stage chi-square test for categorical data.

The usage of this procedure can be seen in several medical studies. For example, Motzer used the O'Brien and Fleming stopping boundaries to end his experiment, entitled "efficacy of everolimus in advanced renal cell carcinoma", early [10]. In addition, Goldberg's study was stopped after 50% of patients responded to his study, entitled "randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer." [11]. Furthermore, Baily used it in their designed trial to examine whether male circumcision was protective against HIV infection [12]. In addition to this, Marcus terminated the Gallium trials earlier by using the O'Brien and Fleming interim analysis. These trials involved follicular lymphoma patients [13].

The O'Brien and Fleming multiple testing procedure has been modified several times by researchers, such as Kung-Jong Lui, who examined the performance of the O'Brien–Fleming multiple testing procedure when intraclass correlations were presented [14]. Along with this, he enhanced the original procedure by increasing the number of stages and the number of treatments. Moreover, in Hammouri's study (2013), the stopping bounds of the O'Brien–Fleming procedure were corrected, and the validation was verified after the correction was applied. When she reviewed the multiple testing methods' stopping bounds, a non-monotonicity problem in the critical values was noticed. The solution was that the number of simulations producing critical values was increased compared to the number of simulations in the initial process. That facilitated the procedure with monotonic critical values. Indeed, when more iterations were added to the simulation, the critical values got larger and made the rejection of the null hypotheses harder, leading to control of the Type I error. Furthermore, the O'Brien–Fleming procedure was updated in Hammouri's work to make it more flexible via three implementations. Each implementation was executed separately. Two of these three implementations are the optimal allocation, where the idea is to allocate more patients to the better treatment after each interim analysis. The other

implementation allows for different sample weights for different stages instead of an equal sample size within the various stages [15].

O'Brien and Fleming (1979) procedure was built using balanced randomization. In this paper, the randomization will be changed to an unbalanced one. Randomization is known to eliminate potential bias and confounders from clinical trials. It is the standard gold method for statistical power. One asset that can be used with multiple-testing methods is unbalanced randomization, which seems favorable due to several constraints. Scientists preferred to use optimal proportions to solve two problems: minimize the number of failures with the power, which is fixed, and maximize the homogeneity test's power with a fixed sample size [16].

Optimal experimental design can be obtained by carefully allocating treatment to study subjects; subjects are randomized to less toxic or more effective treatment regimens. There are many different optimal allocation designs for clinical trials available in the literature. For example, response-adaptive randomization (RAR) design is used to find the optimal allocations for clinical trials with multiple endpoints. RAR designs can be traced to Thompson (1933), Robbins (1952), and Zelen (1969) [17–19]. An example of optimal allocation of patients with RAR designs is the randomized play-the-winner rule. Hu and Rosenberger's (2003) method for optimal RAR procedures involved formalizing the objectives and deriving the optimal allocation proportion for binary responses [20]. Thus, the OWMP will use optimal allocation instead of equal allocation.

One more change will be including the unequal weights in the subsamples. In the literature, Lehmacher and Wassmer (1999) proposed a method that uses adaptive sample size calculations in group sequential trials; the method is for the adaptive planning of sample sizes for group sequential clinical trials [21]. The method was for group sequential trials that combined the results from separate stages using the normal inverse method. The method allows for data-driven reassessments of the sample size without exaggerating the Type I error rate. Next, Proschan and Chi suggested two different two-stage adaptive designs that keep the Type I error rate steady. Proschan's adaptive design is essential to accomplish an anticipated statistical power while restraining the maximum sample size. Furthermore, Chi's adaptive design consists of the main stage with adequate power to reject the null hypothesis and an implementation stage that permits increasing the sample size if the actual effect size is smaller than anticipated [22].

Usually, when a new method is developed from previous methods, a Monte Carlo simulation is used to validate the new one. A Monte Carlo simulation is a system for doing what-if analysis that allows users to measure the reliability of different analyses' results and inferences. In the 1940s, Jon von Neumann and Ulam developed the Monte Carlo simulation, a handy statistical tool for evaluating multiple scenarios in-depth to study uncertain situations. Additionally, simulation studies are associated with pseudo-random sampling, which creates data from computer experiments. Since data generation processes are known in advance, simulation studies have the advantage of understanding and studying the performance of statistical methods [23–25].

In this current study, a new method has been developed that incorporates optimal allocation and varying sample weights at different stages, together with the O'Brien and Fleming multiple testing procedure named the optimal weighted multiple-testing procedure. Furthermore, Type I error and power have been studied to determine if the new method is effective using Monte Carlo simulations.

2. Materials and Methods

The process outlined by O'Brien and Fleming will be reviewed first, since it is the base for this work. This procedure can be used in any clinical trial comparing two treatments. The response to each treatment must be measured using a binary outcome where treatments' outcomes should be collected independently.

2.1. The O'Brien and Fleming Procedure

A description of the steps for the O'Brien and Fleming procedure is presented here, along with an explanation of how critical values were calculated for controlling Type I errors [10]:

2.1.1. Statement of the O'Brien and Fleming Procedure

The data are reviewed and tested periodically, with n_1 subjects receiving treatment 1 and n_2 subjects receiving treatment 2 for each stage. Since there are K stages, the maximum number of subjects is $N = K(n_1 + n_2)$. Where K can take values from one to five. All of these values are fixed in advance.

Initially, n_1 and n_2 subjects are assigned to treatment 1 and treatment 2, respectively.

If $\frac{1}{K}\chi^2_{(1)} \geq P(K, \alpha)$, then the experiment is terminated, and the hypothesis of having a difference between treatments (H_a) is accepted. Where $\chi^2_{(1)}$ is the usual Pearson chi-square, α is the size of the test, and $P(K, \alpha)$ is a critical value obtained from an O'Brien and Fleming table (Table 1). Otherwise, if the critical value is not exceeded, the next $n_1 + n_2$ subjects are randomized, and their measurements are observed.

Table 1. The O'Brien and Fleming original critical values.

α	Number of Stages (K)				
	1	**2**	**3**	**4**	**5**
0.5	0.462	0.656	0.750	0.785	0.819
0.1	2.67	2.859	2.907	2.979	3.087
0.09	2.866	3.031	3.073	3.147	3.283
0.08	3.077	3.197	3.24	3.338	3.467
0.07	3.294	3.363	3.437	3.546	3.663
0.06	3.576	3.652	3.683	3.853	3.889
0.05	3.869	3.928	3.940	4.170	4.149
0.04	4.289	4.231	4.264	4.477	4.584
0.03	4.800	4.722	4.700	4.964	5.045
0.02	5.490	5.392	5.462	5.555	5.789
0.01	6.667	6.574	6.503	6.864	6.838
0.005	7.885	7.818	7.442	7.890	8.037
0.001	10.062	10.240	10.202	11.060	10.600

In general, for the ith test, the study is terminated, and the hypothesis of having a difference between treatments (H_a) is accepted if $\frac{i}{K}\chi^2_{(i)} \geq P(K, \alpha)$ where $\chi^2_{(i)}$ is the usual Pearson chi-square based on all data collected up to the ith test. If not, the data for next stage will be collected.

If, after completing K tests, $\chi^2_{(K)}$ does not exceed $P(K, \alpha)$, the study is terminated with the conclusion that the hypothesis of having a difference between treatments (H_a) is rejected at the α level of significance.

2.1.2. Evaluation of the Stopping Bound

To evaluate the stopping bound, for each $i = 1, \ldots, K$, and $j = 1, 2$, let π_j represents the success rate for treatment j and let y_{ji} be the number of successes with treatment j occurring after the $(i-1)$st test but prior to the ith test. Moreover, define $S_{ji} = \sum_{k=1}^{i} y_{jk}$, and $p_{ji} = \frac{S_{ji}}{in_j}$. The hypothesis being tested H_0 is that $\pi_1 = \pi_2 = \pi$, where π is an unknown constant between 0 and 1. With this notation, $\chi^2_{(i)}$ maybe presented as $\chi^2_{(i)} = Z_i^2$, where $Z_i = \frac{P_{1i} - P_{2i}}{\hat{V}ar(P_{1i} - P_{2i})^{\frac{1}{2}}}$. Such that:

$$\hat{V}ar(P_{1i} - P_{2i}) = \frac{S_{1i} + S_{2i}}{i^2 n_1 n_2}\left(1 - \frac{S_{1i} + S_{2i}}{i(n_{1+}n_2)}\right), \quad (1)$$

is the pooled estimate of variance. An approximate expression for Z_i, obtained by replacing the estimate of variance with its expectation is given by:

$$Z_i^* = \sum_{k=1}^{i} \frac{U_k}{\sqrt{k}} \quad (2)$$

where $U_i \sim NID(0, 1), i = 1, \ldots, K$. Thus approximate values for $P(K, \alpha)$ may be obtained by generating standard normal variates $(U_1, \ldots, U_K)$ and evaluating the percentiles of $max\{T_i\}, 1 < i < K$, where $T_i = \frac{\left(\sum_{k=1}^{i} U_k\right)^2}{K}$. Note that $max\{T_i\}$ has the same distribution as $max\{[W(i/K)]^2\}, 1 < i < K$, where $\{W(t)|0 < t < 1\}$ represents Brownian motion. This algorithm is presented graphically in Figure 1.

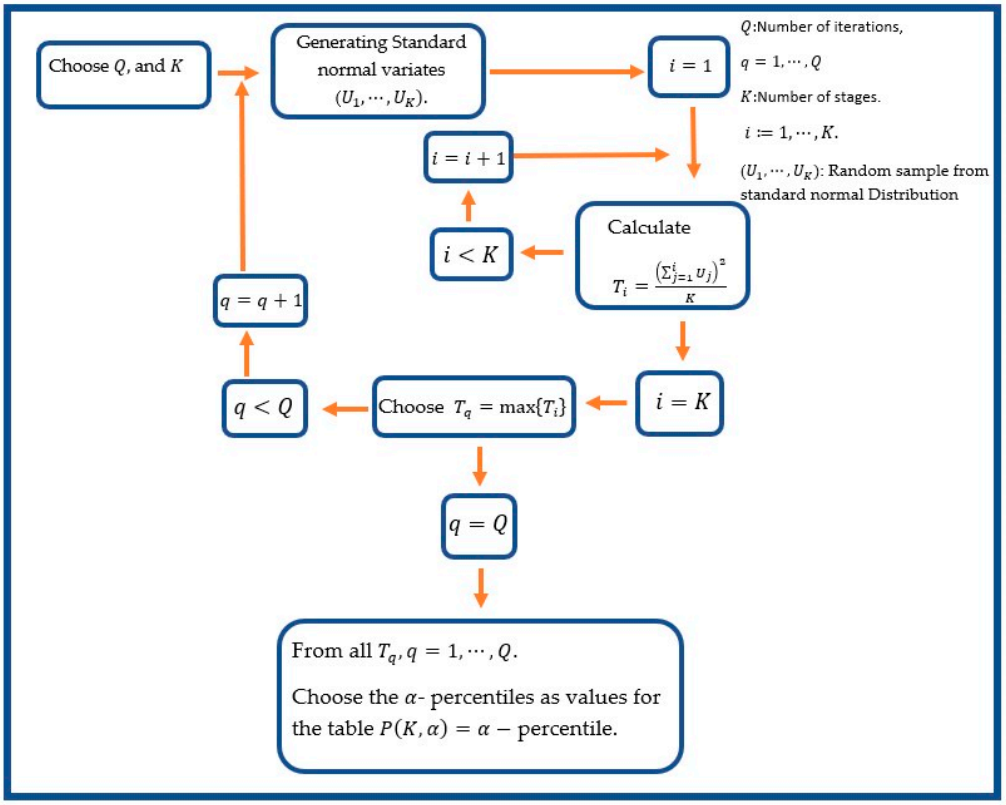

Figure 1. Evaluation of the O'Brien and Fleming Stopping Bound.

Percentile estimates based on 10,000 samples are listed in Table 1 [16].

As a result of the number of iterations used to generate the O'Brien and Fleming critical values, they did not exhibit monotonicity behavior. Based on the original algorithm explained earlier in the methodology, Hammouri has used one million iterations instead of 10,000 to calculate and correct the critical values and found the following values in Table 2 [15]:

Table 2. Corrected O'Brien–Fleming critical values.

α	Number of Stages (K)				
	1	**2**	**3**	**4**	**5**
0.5	0.4547	0.6546	0.7439	0.8013	0.8431
0.1	2.7042	2.8195	2.9247	3.0047	3.0650
0.09	2.8730	2.9817	3.0877	3.1668	3.2275
0.08	3.0633	3.1646	3.2700	3.3498	3.4114
0.07	3.2814	3.3754	3.4799	3.5594	3.6220
0.06	3.5348	3.6207	3.7251	3.8034	3.8669

Table 2. *Cont.*

α	Number of Stages (K)				
	1	**2**	**3**	**4**	**5**
0.05	3.8399	3.9152	4.0191	4.0961	4.1602
0.04	4.2177	4.2809	4.3836	4.4599	4.5243
0.03	4.7099	4.7622	4.8587	4.9341	5.0008
0.02	5.4106	5.4537	5.5396	5.6148	5.6827
0.01	6.6393	6.6618	6.7353	6.8021	6.8764
0.005	7.8863	7.9019	7.9529	8.0094	8.0803
0.001	10.8280	10.8527	10.8618	10.9263	10.9820

2.2. Optimal Allocation

Optimal allocation aims to use better resources by allocating more subjects to superior treatment. In the interim analyses with binary responses, the optimal allocation is widely used. For the OWMP, the new allocation ratio for both groups will depend on the optimality criteria, based on the success counts from the previous stage in the interim analysis so that the allocation might change from stage to stage of the study.

Accordingly, optimal allocation is illustrated as follows:

Let $f(P_A,\ P_B)$ be a function to compare two binomial probabilities. Three usually used functions are the simple difference $= P_A - P_B$, the relative risk $= \frac{P_B}{P_A}$ and the odds ratio $= \frac{P_A q_B}{P_B q_A}$. These functions are estimated by replacing P_A by $\hat{P}_A$ and P_B by $\hat{P}_B$, where $\hat{P}_A$ and $\hat{P}_B$ are the proportions of successes observed for group A and group B, respectively. The delta method can be used to estimate the asymptotic variances of these estimators a Var $\{f(P_A,\ P_B)\}$.

The idea is to find the optimal allocation $R = \frac{n_A}{n_B}$ that has a fixed asymptotic variance to minimize the expected number of failures by allocating the patients to better treatment. The unknown binomial parameters determine the optimal allocation. Rosenberger Stallard, Ivanova, Harper, and Ricks needed to develop a sequential design to estimate the optimal design. Let $X_1,\ \ldots,\ X_n$ be a binary response (with two values, success = 1 and failure = 0), and $T_1,\ \ldots,\ T_n$ are treatment assignment indicators, which have the values of one for treatment A and zero for treatment B. Then, they wrote $N_{A,n} = \sum_{i=1}^{n} T_i$, $N_{B,n} = n - N_{A,n}$, $\hat{P}_{A,n} = \sum_{i=1}^{n} \frac{T_i X_i}{N_{A,n}}$, $\hat{P}_{B,n} = \sum_{i=1}^{n} \frac{(1-T_i)X_i}{N_{B,n}}$, $\hat{q}_{A,n} = 1 - \hat{P}_{A,n}$, and $\hat{q}_{B,n} = 1 - \hat{P}_{B,n}$. Further, they denoted $F_i = \{X_1,\ \ldots,\ X_n, T_1,\ \ldots,\ T_n\}$ and conditional expectation as $E_i(.) = E(.|F_i)$. They stated the following allocation rule:

$$E_{i-1}(T_i) = \frac{\sqrt{\hat{P}_{A,(i-1)}}}{\sqrt{\hat{P}_{A,(i-1)}} + \sqrt{\hat{P}_{B,(i-1)}}} \tag{3}$$

Therefore, this rule substitutes the unknown success probabilities in the optimal allocation rule with the current estimate of the proportion of successes on each treatment thus far in the trial. When $P_A, P_B \in (0, 1)$,

$$\frac{N_{A,n}}{n} \ \rightarrow \frac{\sqrt{P_A}}{\sqrt{P_A} + \sqrt{P_B}}\ , \text{ as } n \rightarrow \infty. \tag{4}$$

Then they reached the following formula $\frac{\sqrt{P_A}}{\sqrt{P_A}+\sqrt{P_B}}$ is the optimal allocation [26]. This formula will be used in this current study as $w_i = \frac{\sqrt{P_{(i-1),1}}}{\sqrt{P_{(i-1),1}}+\sqrt{P_{(i-1),2}}}$.

3. The Proposed Procedure OWMP

3.1. The New Methodology

Using the corrected critical values and then combining the procedure with the optimal allocation together with different sample weights, the procedure was enhanced for efficiency.

The method is suggested as follows:

There is a total sample size N, and a number of stages K, as well as the sample weights $\{we_1, \ldots, we_K\}$ and α, which are all chosen in advance. where sample weights $\{we_1, \ldots, we_K\}$ are used to get each stage sample size n_i, where $we_i > 0 \; \forall \; i = 1, \cdots, K$, $we_1 > \ldots > we_K$ and $\sum we_i = 1$. For each K, the sample size for each stage i is calculated as follows:

- For $i = 1, \ldots, K-1$, n_i is calculated as n_i = round $(we_i \times N)$, if n_1 is even. Otherwise n_1 = round $(we_i \times N) + 1$; because equal allocation is used in the first stage. Furthermore, equal allocation is used in other stages when the optimal ratio equals zero or one.
- For the last stage K, $n_K = N - \sum_{k=1}^{K-1} n_k$ to cover the rounding that is used in the previous stages.

Now, for $i = 1$, treatments A and B are assigned to n_{11} and n_{12} subjects, respectively. Where n_{11} and n_{12} must be equal and ($n_1 = n_{11} + n_{12}$). For $i = 2, \ldots, K$, A stage with n_i subjects are divided to n_{i1} and n_{i2} subjects assigned to treatment A and B, respectively.

Where $n_{i1} = round(w_i \times n_i)$ with $w_i = \frac{\sqrt{P_{(i-1),1}}}{\sqrt{P_{(i-1),1}} + \sqrt{P_{(i-1),2}}}$, and $P_{(i-1),1}$ and $P_{(i-1),2}$ are success rates from the previous stage for treatment A, and treatment B, respectively [26]. Therefore, $n_{i2} = (n_i - n_{i1})$ for all $2 \leq i \leq K$. Then for each stage $i = 1, \ldots, K$, subjects are randomized, and their measurements are observed. Each subsample will be added to the previous subsamples for the same treatment. At that time $\frac{i}{K}\chi^2_{(i)}$ is calculated. Where $\chi^2_{(i)}$ is the usual Pearson chi-square statistic. $\frac{i}{K}\chi^2_{(i)}$ is compared to $P(K, \alpha)$, where α is the size of the test, and $P(K, \alpha)$ is a critical value from Table 2. If $\frac{i}{K}\chi^2_{(i)} \geq P(K, \alpha)$, then the study is terminated, and the null hypothesis is rejected. Otherwise $\frac{i}{K}\chi^2_{(i)} < P(K, \alpha)$, and $i = K$, the study is terminated, and the null hypothesis fails to be rejected. However, if $\frac{i}{K}\chi^2_{(i)} < P(K, \alpha)$, and $i < K$, the procedure proceeds to the next stage. The method is illustrated in Figure 2.

3.2. Type I Error and Power to Validate the OWMP

This section utilizes Monte Carlo simulations to investigate the Type I error and the power of the OWMP. A theoretical approach can, in many situations, be challenging to implement, much less to find a precise answer. Using Monte Carlo simulations can provide an alternative to theoretical analysis. In the case of O'Brien and Fleming, they used an approximation distribution, so they used simulation to show that a fixed one-stage chi-square test has the same Type I error rate and power as theirs. So, the same approach was used in the current work, and all simulations were run using SAS software.

3.2.1. Testing Type I Error Algorithm

In order to calculate the Type I error, success probabilities ($P = 0.1, 0.2, 0.3, 0.4, 0.5$), $\alpha = 0.99$ and 0.95 and critical value $P(K, \alpha)$ were chosen with different sample sizes for all values of K. In each case of $K = 1, \ldots, 5$, both subsamples are generated from the same binomial distribution with the same success rate. Assess if the OWMP fails to reject the null hypothesis of no significant difference between groups or accepts that there is a significant difference between groups. The latter result is causing a Type I error. After repeating this 500,000 times, the proportion of rejecting H_0 is calculated and this represents the Type I error (Figure 3).

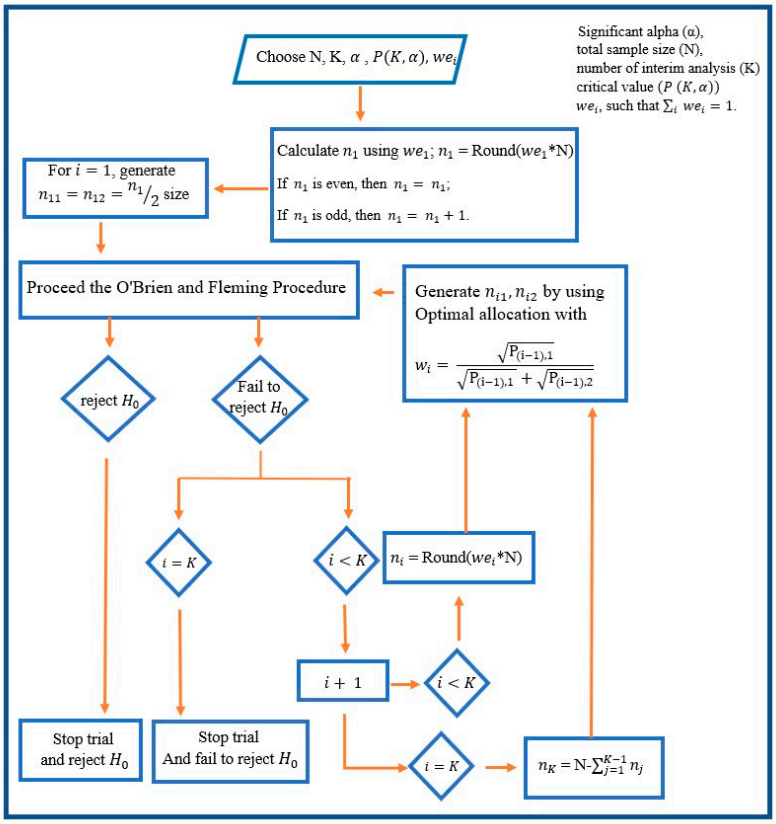

Figure 2. The algorithm for the optimal weighted multiple—testing procedure.

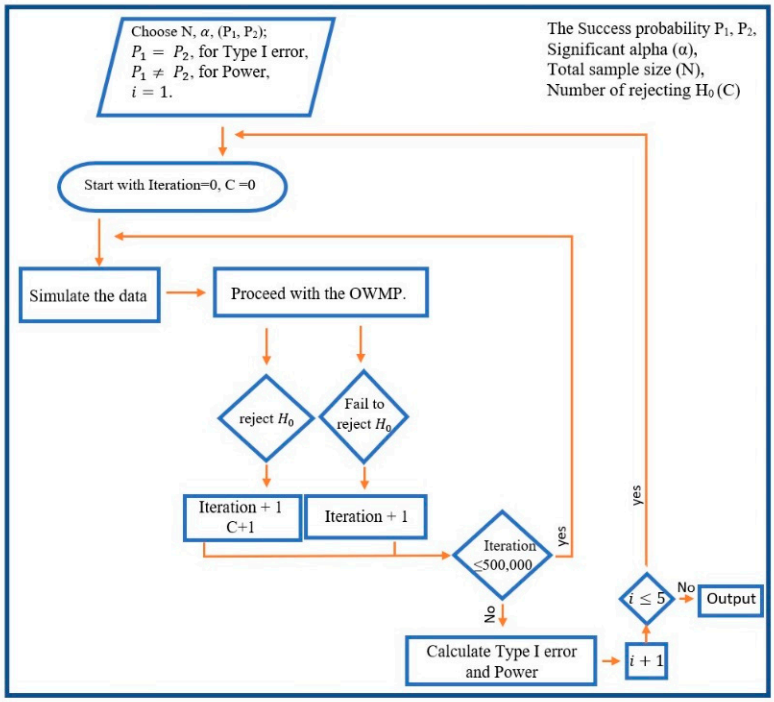

Figure 3. The algorithm to calculate the values of the Type I error and power using Monte Carlo simulations with the OWMP.

3.2.2. Result of Testing Type I Error

Simulations were run to calculate the Type I error for the multiple-testing procedure using SAS code. Sample sizes of various scenarios were considered. Almost identical results were obtained for all sample sizes that were used. Compared to the usual chi-square, the OWMP maintained, if not decreased, the Type I error. With higher K, it was observed that the decreasing trend intensified. The reason is that, with a larger K, the usual chi-square statistic is multiplied by a factor less than one and it is compared with a larger critical value. So, rejecting H_0 becomes harder. For illustration, results for sample size 250 with $\alpha = 0.05$ and sample size of 300 with $\alpha = 0.01$ were reported in Tables 3 and 4, respectively.

Table 3. Type I values resulted from Monte Carlo simulations for α = 0.05 with the OWMP.

		Number of Stages (*K*)				
		1	**2**	**3**	**4**	**5**
P	0.1	0.0503	0.0486	0.0458	0.0442	0.0415
	0.2	0.0499	0.0487	0.0478	0.0457	0.0430
	0.3	0.0499	0.0488	0.0476	0.0461	0.0433
	0.4	0.0487	0.0485	0.0477	0.0463	0.0438
	0.5	0.0499	0.0500	0.0497	0.0453	0.0422

Table 4. Type I values resulted from Monte Carlo simulations for α = 0.01 with the OWMP.

		Number of Stages (*K*)				
		1	**2**	**3**	**4**	**5**
P	0.1	0.0093	0.0088	0.0082	0.0079	0.0075
	0.2	0.0098	0.0098	0.0094	0.0089	0.0084
	0.3	0.0101	0.0100	0.0094	0.0091	0.0088
	0.4	0.0104	0.0100	0.0096	0.0092	0.0089
	0.5	0.0094	0.0099	0.0097	0.0094	0.0090

In the first case, with $\alpha = 0.05$ and a sample size of 250, when $K = 1$, the values for the Type I error ranged between 0.0499 and 0.0503. While the number of K increased, the values of the Type I error monotonically decreased, where the values are between 0.0415 and 0.0438 at $K = 5$, which is less than 0.05, which is an even more acceptable error than the usual chi-square procedure.

Likewise, various sample sizes were used to compute the Type I error values resulting in the same conclusion. For example, the Type I error values with a sample size of 80 and 580 were calculated. The values ranged between 0.0418 and 0.0506, 0.0438 and 0.0507, respectively, for $\alpha = 0.05$, and all values of K.

Similarly, Type I error values displayed a monotonic behavior with $\alpha = 0.01$, since the values ranged between 0.0093 and 0.0104, in the first case, with $K = 1$. A decrease in Type I errors has also been noted, while a rise in K values has been applied, which is satisfactory since the errors do not exceed 0.0104.

Furthermore, other samples were calculated in order to determine the Type I error values. For example, values for $\alpha = 0.01$ and sample sizes of 80 and 630 were calculated. It has been observed that the error values ranged from 0.0090 to 0.0101 and from 0.0084 to 0.0101, respectively, indicating that the OWMP is working effectively regarding Type I errors.

3.2.3. Testing Power Algorithm

To evaluate the power values, a probability value $P_1 = 0.1$ was chosen for all cases and a different success probability P_2 was chosen from the set {0.15, 0.2, 0.25, 0.3} with $\alpha = 0.01$ and $\alpha = 0.05$, the number of interim analyses being 500,000, and the corrected O'Brien and Fleming critical values $P(K, \alpha)$. Sample sizes were chosen such that the sample guaranteed the power values to be equal to 0.8 using the usual chi-square test power calculation.

In each case of $K = 1, \ldots, 5$, the subsamples are generated from two different binomial distributions with different P_1 and P_2 to make sure that the alternative hypothesis is true. Then, the OWMP was used to examine if H_a (that there is a difference between the two groups) is rejected, or a significant difference is found and H_a is accepted. After repeating this 500,000 times, the proportion of accepting H_a is computed, and this is the power rate (because H_a was assured to be true). The process of computing the power is illustrated in Figure 3.

3.2.4. Result of Testing Power

The OWMP was able to preserve acceptable power values with the new implementation. The OWMP was studied for $\alpha = 0.05$ and 0.01, and several values of N that guaranteed a power of 0.8, and SAS code was used to calculate the power for each case with a variety of K values. Tables 5 and 6, Figures 4 and 5 show the power values from the 500,000 simulations.

Table 5. Power values for the OWMP with $\alpha = 0.05$, $P_1 = 0.1$, and $P_2 = 0.15,\ 0.2,\ 0.25,$ and 0.3.

P_1	Number of Stages (K)						
	P_2	n	1	2	3	4	5
0.1	0.15	1366	0.8020	0.7959	0.7890	0.7857	0.7822
	0.2	394	0.8046	0.7977	0.7917	0.7858	0.7786
	0.25	200	0.8164	0.8062	0.8008	0.7953	0.7892
	0.3	120	0.8133	0.8013	0.7931	0.7795	0.7726

Table 6. Power values for the OWMP with $\alpha = 0.01$, $P_1 = 0.1$, and $P_2 = 0.15, 0.2, 0.25, 0.3$.

P_1	Number of Stages (K)						
	P_2	N	1	2	3	4	5
0.1	0.15	2032	0.8010	0.7994	0.7955	0.7918	0.7875
	0.2	588	0.8037	0.8016	0.7943	0.7897	0.7840
	0.25	296	0.8113	0.8079	0.8002	0.7941	0.7888
	0.3	182	0.8085	0.8057	0.8006	0.7926	0.7841

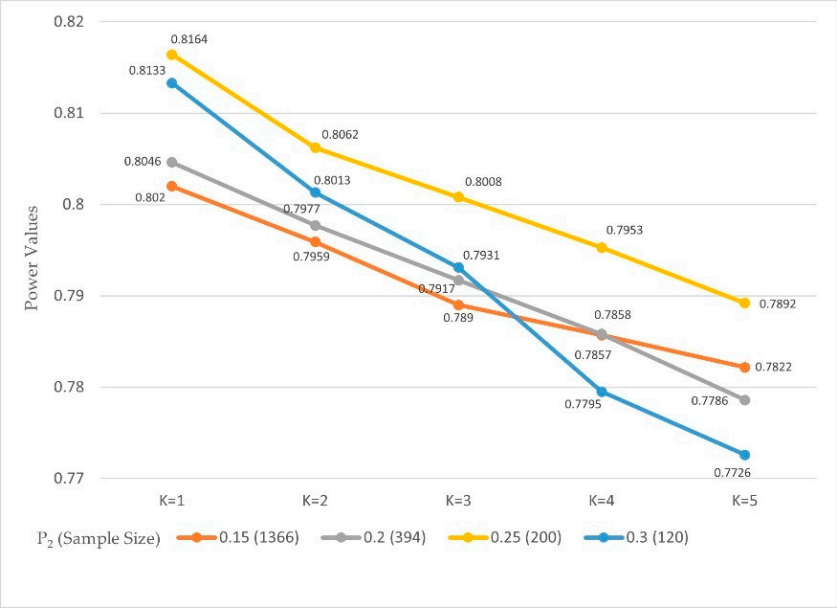

Figure 4. Power values of OWMP with $\alpha = 0.05$, $P_1 = 0.1$, and $P_2 = 0.15, 0.2, 0.25, 0.3$.

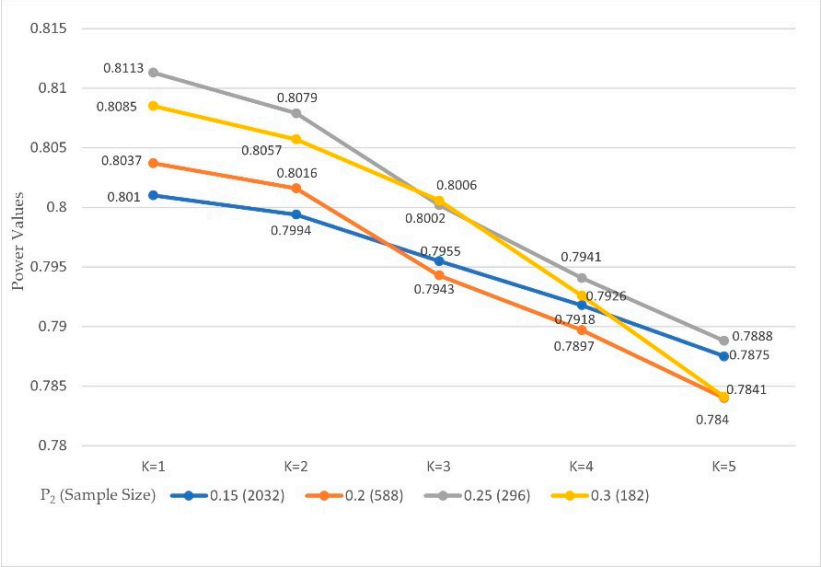

Figure 5. Power values for OWMP with $\alpha = 0.01$, $P_1 = 0.1$, and $P_2 = 0.15, 0.2, 0.25, 0.3$.

The probability value $P_1 = 0.1$ was fixed, and the power rates with various values for $P_2 = 0.15,\ 0.2,\ 0.25$, and 0.3, were studied.

The sample sizes 1366, 396, 200, and 120 for $\alpha = 0.05$, and 2032, 588, 292, and 182 for $\alpha = 0.01$ were used to ensure that the power values were 0.8.

It was noticed that with $\alpha = 0.05$, the results for the power values with $K = 1$ were between 0.8046 and 0.8164. Then, the values showed a decreased behavior when the values of K were increased because the power values, when $K = 5$, were between 0.7726 and 0.7892, with marginal errors not more than 0.0274 between the power values and the 0.8.

The power values with $K = 1$ and $\alpha = 0.01$, were between 0.8010 and 0.8113. It was further found that by comparing the power results with the values of K, the power values monotonically decreased as K values increased. However, the power values when $K = 5$ were between 0.7840 and 0.7888 with marginal errors not more than 0.0160.

In both α values, the marginal errors are negligible to be considered.

3.3. Calculating Rejection Rates for Each Stage

In this section, for each stage, the null hypothesis rejection rates, along with the sample sizes required to reject it were calculated.

3.3.1. Calculating Rejection Rates for Each Stage When H_a Is True, and the Difference Is Presented

The rejections were calculated with a standard power value of 0.8, with the probabilities of success of $P_1 = 0.1$ and $P_2 = 0.2$ with $\alpha = 0.01$, and the sample size was 588. Table 7 and Figure 6 demonstrate the needed sample size and the number of rejections of H_0 occurring at stage i with 500,000 iterations.

Table 7. Sample sizes and percentages of rejections of H_0 occurring at stage i with 500,000 iterations. Where $P_1 = 0.1$ and $P_2 = 0.2$ with $\alpha = 0.01$, and the sample size equals 588.

		Number of Stages (*K*)			
		2	**3**	**4**	**5**
i	First Stage	488(35%)	338 (3%)	300(0%)	180 (0%)
	Second Stage	588 (65%)	488 (55%)	430 (25%)	330 (4%)
	Third Stage		588 (42%)	530 (49%)	450 (35%)
	Fourth Stage			588 (26%)	530 (38%)
	Fifth Stage				588 (23%)

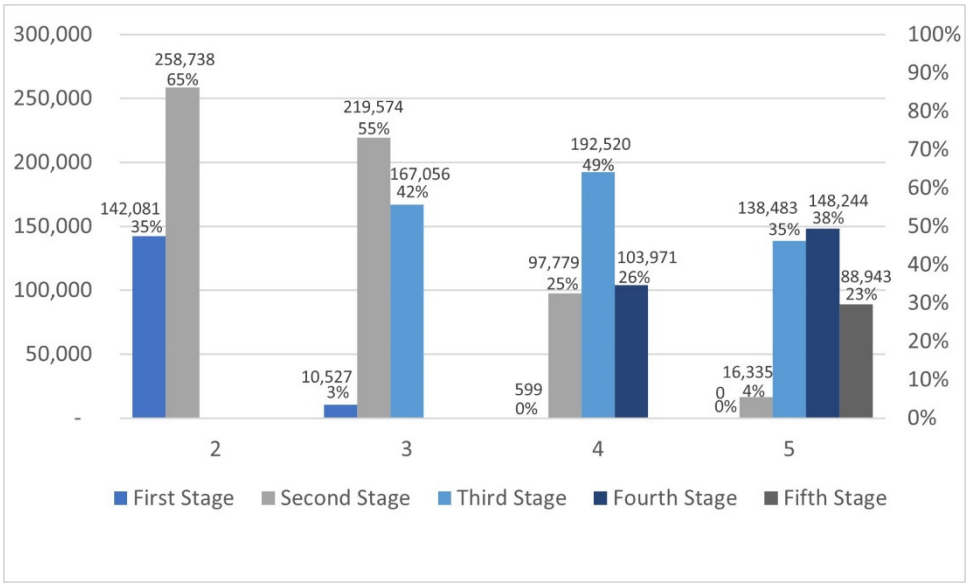

Figure 6. Percent of rejections of H_0 occurring at stage i with 500,000 iterations.

With $K = 2.35\%$ of rejections occurred in the first stage, and 65% occurred in the second stage, so the whole sample size was not needed to reject the H_0 hypothesis in more than a third of the cases. Based on $K = 3$, the sample size needed to reject the H_0 hypothesis in the second stage was 488, which resulted in a 55% rejection rate. The highest rejection rate with $K = 4$ was in the third stage with 49%, and the needed sample size was 530 to reject H_0.

In addition, in 74% of the cases, the whole sample was not needed. At $K = 5$, the highest rejection percentage was 38%, with a 530 sample size, and the highest percentage was at the fourth stage. Rejection occurred in 77% of cases earlier in the process.

3.3.2. Calculating Rejection Rates for Each Stage When H_0 Is True, and the Difference Is Not Presented

The rejections were calculated with $\alpha = 0.05$, and the sample size was 394. Based on 500,000 iterations, Table 8 and Figure 7 below illustrate the required sample sizes and the number of rejections (percentages) at each stage for all values of K. It needs to be noted that these percentages are out of the 5% rejecting rate. The decision rules for this multiple-testing procedure are nearly identical to the usual chi-square one-stage procedure in the absence of early termination when H_0 is true.

Table 8. Based on 5% of the 500,000 iterations, values for sample sizes and acceptance rates at stage i when H_0 is true.

		Number of Stages (*K*)			
		2	**3**	**4**	**5**
i	First Stage	250 (1%)	220 (0%)	200 (0%)	160 (0%)
	Second Stage	300 (99%)	270 (12%)	250 (1%)	210 (0%)
	Third Stage		300 (88%)	280 (23%)	250 (4%)
	Fourth Stage			300 (76%)	280 (26%)
	Fifth Stage				300 (70%)

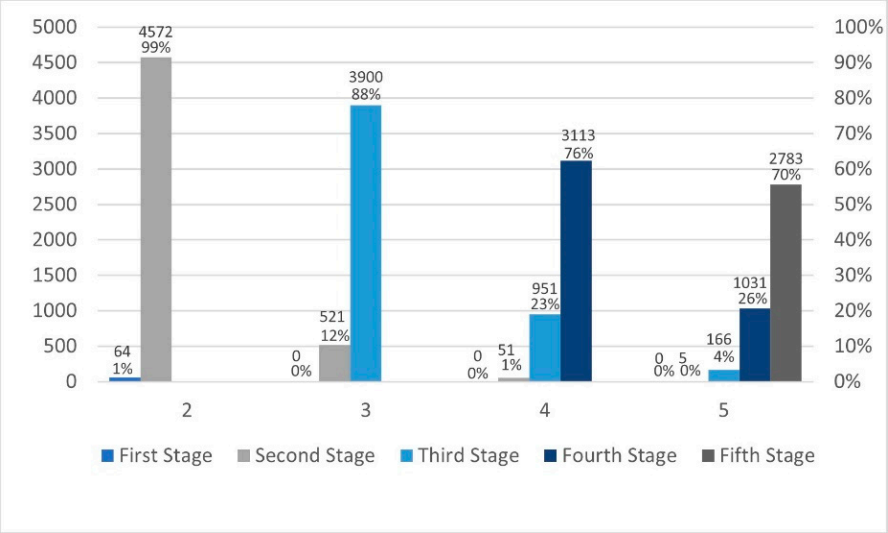

Figure 7. Percent of accepting H_0 occurring at stage i with 500,000 iterations.

4. Examples

4.1. Example 1: Real Life Example

Three hundred individuals' behavior was examined regarding their parents' smoking status and how it influences their behavior. Participants were chosen based on their smoking status: 150 smokers and 150 non-smokers, then their parents' smoking status were recorded. After assigning data for each value of K from 1 to 5, OWMP was applied. The necessary sample size was also noted when the hypothesis was rejected. For illustrative purposes, both methods were applied when $K = 4$: the original O'Brien and Fleming method, as well as the OWMP.

Briefly, the procedure is explained for $K = 4$: $we_1 = 40\%$, $we_2 = 25\%$, $we_3 = 20\%$, $we_4 = 15\%$, and by using the weighted formula, we got: $n_1 = 120$, $n_2 = 76$, $n_3 = 60$, $n_4 = 44$.

For the first subsample: n_{11} and n_{12} equal 60 because the subsample was divided equally. The chi-square statistic equals 1.534 after multiplying it by one fourth, which is not greater than the critical value of 4.0961, so we failed to reject H_0.

For the second stage, with $n_2 = 76$, n_{21} and n_{22} are needed to be recalculated by using the optimal allocation. Where $n_{21} = round(w_2 \times 76)$ with $w_i = \frac{\sqrt{19}}{\sqrt{19}+\sqrt{13}}$.

The optimal allocation resulted in $n_{21} = 42$ and $n_{22} = 34$, with the chi-square statistic equal to 23.05, which is larger than the critical value after multiplying it by 2/4. Thus, H_0 is rejected.

Only two stages out of four stages, with only 196 out of the 300 participants, were needed to end the experiment and get a significant difference between the two treatments. In addition, using the original O'Brien and Fleming method, the following results were observed:

For the original procedure, equal subsamples are used as follows: $n_1 = n_2 = n_3 = n_4 = 76$

For the first stage, with $n_{11} = n_{12} = 38$, the chi-square statistic equals 0.163. Multiplying it by one-fourth resulted in a value that is not greater than the critical value of 4.096, so a significant difference was not found.

For the second stage, with $n_{21} = n_{22} = 38$, and the chi-square statistic for the cumulative data equals 5.856. Multiplying it by two-fourths resulted in a value that is not greater than the critical value, so a significant difference was not found again.

For the third stage, with $n_{31} = n_{32} = 38$, the chi-square statistic for the cumulative data equals 39.097. Multiplying it by third-fourth results in a value that is greater than the critical value, so a significant difference was found with using 228 participants in three

stages, compared to 196 with two stages by using OWMP, which means OWMP reached the same conclusion with fewer stages and participants than the original procedure.

Table 9 was recapped in Figure 8 to summarize the results.

Table 9. The results of the OWMP for a real-life example.

i	Critical Values	n_1	n_2	X_1	X_2	Total Sample Size	Usual Chi-Square	$\overset{i}{K}\chi^2_{(i)}$
Result of Case 1 ($K = 1$) ($we_1 = 100\%$)								
1	3.84	150	150	91	14	300	86.9	86.9
Result of Case 2 ($K = 2$) ($we_1 = 60\%$, $we_2 = 40\%$)								
1	3.92	90	90	37	13	180	15.95	7.975
Result of Case 3 ($K = 3$) ($we_1 = 40\%$, $we_2 = 35\%$, $we_3 = 25\%$)								
1	4.02	60	60	19	13	120	1.534	0.5113
2	4.02	118	108	63	13	226	41.201	27.467
Result of Case 4 ($K = 4$) ($we_1 = 40\%$, $we_2 = 25\%$, $we_3 = 20\%$, $we_4 = 15\%$)								
1	4.02	60	60	19	13	120	1.534	0.3835
2	4.02	102	94	47	13	196	25.951	12.976
Result of Case 5 ($K = 5$) ($we_1 = 30\%$, $we_2 = 25\%$, $we_3 = 20\%$, $we_4 = 15\%$, $we_5 = 10\%$)								
1	4.16	45	45	13	11	90	0.227	0.0454
2	4.16	85	81	32	13	166	9.791	0.0226
3	4.16	121	105	64	13	226	41.074	24.644

X_1 : success probability of treatment A, X_2 : success probability of treatment B.

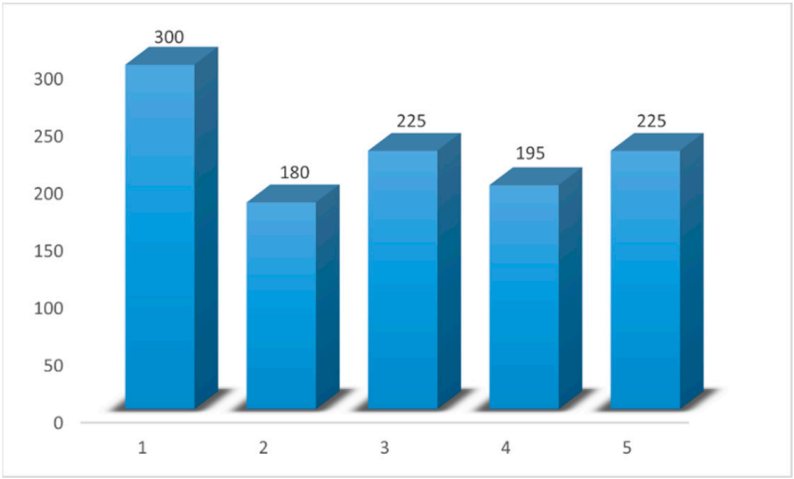

Figure 8. Sizes of samples necessary to reach the rejection of H_0 for the real-life example.

We can see a difference with only 180 patients, which is less than the 300 patients that would have been used in a conventional clinical trial.

4.2. Example 2: Computational Example

Simulated data of 400 subjects for two groups were used, with a success rate of 0.23 and 0.4 for group 1 and group 2, respectively (consequently, the alternative hypothesis is correct). Data from this trial were simulated, and the case when K equals three ($K = 3$) was studied in detail, where we_1 = 45%, we_2 = 35%, we_3 = 20%, and by using the weighted formula, $n_1 = 180,\ n_2 = 140,$ and $n_3 = 80$.

For the first subsample: $n_{11} = n_{12} = 90$, because the subsample was divided equally into the two groups in the first step. The chi-square statistic equals 0.88 after multiplying it by one-third. The result is not greater than the critical value of 4.02, so H_0 failed to be rejected.

For the second stage, with $n_2 = 140$, n_{21} and n_{22} were calculated by using the optimal allocation. That resulted in $n_{21} = 63$ and $n_{22} = 77$. Moreover, the chi-square statistic equals 5.42, which is larger than the critical value after multiplying it by two thirds. Thus, H_0 was rejected, and a significant difference was found.

In this case, the trial is terminated using only two stages out of three stages, with only 320 of the 400 subjects needed to get a significant difference.

The remainder of the results are shown in Table 9.

Table 10 was recapped in Figure 9 to summarize the results.

Table 10. The results of the OWMP for simulated example.

i	Critical Values	n_1	n_2	X_1	X_2	Total Sample Size	Chi-Square	$\frac{i}{K}\chi^2_{(i)}$
Result of Case 1 ($K = 1$) ($we_1 = 100\%$)								
1	3.84	200	200	46	79	400	12.7	12.7
Result of Case 2 ($K = 2$) ($we_1 = 70\%$, $we_2 = 30\%$)								
1	3.92	140	140	32	53	280	7.508	3.73
2	3.92	192	208	43	82	400	13.473	13.473
Result of Case 3 ($K = 3$) ($we_1 = 45\%$, $we_2 = 35\%$, $we_3 = 20\%$)								
1	4.02	90	90	22	32	180	2.646	0.88
2	4.02	153	167	36	64	320	8.134	5.42
Result of Case 4 ($K = 4$) ($we_1 = 40\%$, $we_2 = 25\%$, $we_3 = 20\%$, $we_4 = 15\%$)								
1	4.02	80	80	19	29	160	2.976	0.74
2	4.02	125	135	29	51	260	6.475	3.23
3	4.02	164	176	38	68	340	9.431	7.07
Result of Case 5 ($K = 5$) ($we_1 = 30\%$, $we_2 = 25\%$, $we_3 = 20\%$, $we_4 = 15\%$, $we_5 = 10\%$)								
1	4.16	60	60	16	21	120	0.977	0.19
2	4.16	107	113	25	41	220	4.368	1.74
3	4.16	143	157	32	59	300	8.184	4.91

X_1 : success of probability of treatment A, X_2 : success of probability of treatment B.

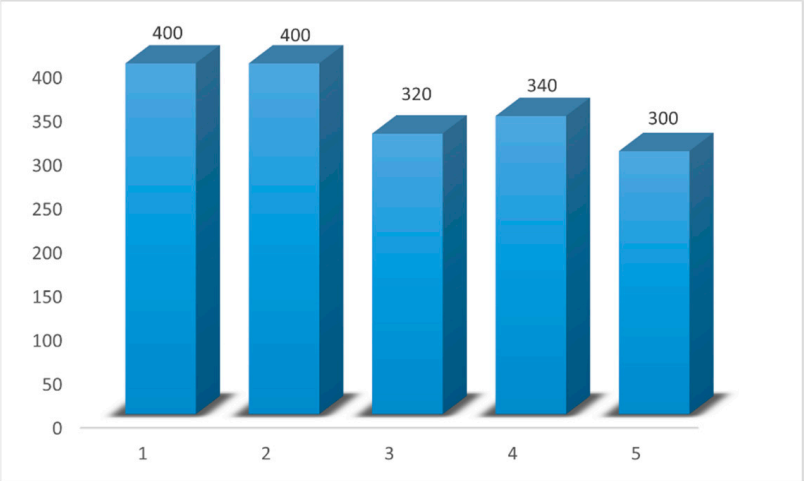

Figure 9. Sizes of samples necessary to reach the rejection of H_0 for the simulated example.

5. Discussion

Group sequential procedures based on multiple primary critical points are used as central cornerstones of strata tests to increase the efficiency of clinical trials. These multiple critical endpoints provide a complete characterization of the effect of any intervention in trials. Several of these procedures have been proposed. O'Brien and Fleming (1979) first proposed a proper multiple testing procedure. Its original critical points were modified for more accuracy and to exhibit a monotonic behavior with 1,000,000 iterations, by Hammouri [15].

As a generalization, in this work, the O'Brien and Fleming procedure was combined with two different modifications for more fixable trials: the optimal allocation and unequal weighted allocation for subsamples. These two allocations were chosen since the optimal allocation implementation assigns more subjects to the more effective treatment. At the same time, the unequal weighted allocation allows for different subsample weights for different stages instead of equal allocation at each phase. Thus, it is possible to terminate the trial early and orient the largest number of participants toward the most effective treatment. The Type I error and power was calculated for several scenarios with several sample sizes to check the validity of this work. It was noticed that the new combination decreased the values of the Type I error and maintained the power in comparison with the usual chi-square, which indicates that OWMP is effective.

In detail, the Type I error values were computed with various α values and sample sizes. The first case was with $\alpha = 0.05$, and the sample size equals 250. We noticed that the initial values for the Type I error with $K = 1$ were between 0.0499 and 0.0503. The critical values monotonically decreased, whereas the number of K increased, such that when the $K = 5$, the Type I error values were between 0.0415 and 0.0438, which is less than 0.05, which means it is an acceptable error, and better than the error in the original procedures.

Moreover, with $\alpha = 0.05$, and the sample sizes equal to 80 and 580, the Type I error values were between 0.0418 and 0.0506, 0.0438 and 0.0507, respectively, which is acceptable since the values are not more than 0.0507.

Similarly, with $\alpha = 0.01$ the values of the Type I error values, in general, took on an analogous monotonic behavior, and the values were between 0.0084 and 0.0104. Again, the values decreased while the K value increased, which is acceptable since the values are still not more than 0.0104. It is worth mentioning that the higher values occurred for $K = 1$, which represents the usual chi-square. Alternatively, all $K = 2, \ldots, 5$ values were less than 0.05.

To determine whether the proposed procedure maintains the acceptance rate of the H_a hypothesis when it is true (power), several comparisons were made with different values for success rates and sample sizes. The probability values were 0.1 with 0.15, 0.2,

0.25, and 0.3 with sample sizes 1366, 750, 200, and 120, for $\alpha = 0.05$, and 2032, 588, 292, and 182, for $\alpha = 0.01$. The power results indicated that OWMP preserved the acceptance rate (power) despite the slightly decreasing values. The values decreased according to the division of the chi-square values by the number of stages K in the interim analysis. This decrease in the chi-square values makes the rejection of H_0 harder. Nevertheless, values are still acceptable since the power values were between 0.8020 and 0.8164 for $K = 1$, and between 0.7726 and 0.7892 for $K = 5$. Thus, the difference between the values is not more than 0.044 for $\alpha = 0.05$. In addition, for $\alpha = 0.01$, the power values are acceptable since they remain between 0.8010 and 0.8113 for $K = 1$, and between 0.7840 and 0.7888 for $K = 5$. Thus, the difference between the values is not more than 0.0273.

Furthermore, the rejected iteration numbers and percentages were calculated under two scenarios of H_a being true or false. Under H_a being false, across the 500,000 samples, there was no significant difference between the OWMP and the usual chi-square test, and in the absence of early termination, the procedure's decision rules are almost identical to those of the chi-square test. In contrast, when H_a is true, the OWMP terminated the trial early in most cases, indicating that a smaller sample size is required to reject the H_a hypothesis compared to the usual chi-square test.

The usual chi-square procedure collects data in much the same way as the O'Brien and Fleming multiple testing procedure and the proposed OWMP in this paper, but at once. Because the chances of a Type I error and power are basically the same for all procedures, the multiple testing procedures appear to gain nearly a considerable advantage over the usual chi-square procedure. Furthermore, the new OWMP is more flexible than the original multiple testing procedure. Researchers who would have otherwise adopted a single sampling design can now review their data periodically and terminate the study early if one treatment proves to be superior to the other, without sacrificing any of the advantages of sequential methods.

As a result, the OWMP proposed in this work is believed to be more effective than both the single sample approach and the O'Brien and Fleming multiple testing.

In our subsequent work, a plan was made to compare the OWMP performance to other previous procedures developed for testing binary outcomes, and to explore the possibility of generalizing other multiple testing procedures by combining different allocations to prechosen approaches [27–34]. We also plan to modify the original O'Brien and Fleming procedure by combining it with several new allocation methods. Furthermore, the proposed approach will be further expanded, by using the modification used for the O'Brien and Fleming original procedure, from two to ten treatments and the number of stages from five to ten by using more iterations to broaden our scope [35].

Author Contributions: Formal analysis, H.H. and R.A.M.; Methodology, H.H. and J.A.; Project administration, H.H.; Software, H.H., R.A.M. and J.A.; Supervision, H.H.; Validation, H.H.; Writing—original draft, H.H., M.A., R.A.M. and J.A.; Writing—review & editing, H.H., M.A. and R.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be obtained from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Glancey, J.L.; Upadhyaya, S.K. An improved technique for agricultural implement draught analysis. *Soil Tillage Res.* **1995**, *35*, 175–182. [CrossRef]
2. Chow, S.-C.; Chang, M. Adaptive design methods in clinical trials—A review. *Orphanet J. Rare Dis.* **2008**, *3*, 11. [CrossRef] [PubMed]
3. Jennison, C.; Turnbull, B.W. *Group Sequential Methods with Applications to Clinical Trials*; CRC Press: Boca Raton, FL, USA, 1999. [CrossRef]
4. Bretz, F.; Maurer, W.; Brannath, W.; Posch, M. A graphical approach to sequentially rejective multiple test procedures. *Stat. Med.* **2009**, *28*, 586–604. [CrossRef] [PubMed]
5. Fu, Y. Step-Down Parametric Procedures for Testing Correlated Endpoints in a Group-Sequential Trial. *Stat. Biopharm. Res.* **2018**, *10*, 18–25. [CrossRef]
6. Shun, Z. Stopping Boundaries Adjusted for Sample Size Reestimation and Negative Stop. *J. Biopharm. Stat.* **2002**, *12*, 485–502. [CrossRef]
7. Urach, S.; Posch, M. Multi-arm group sequential designs with a simultaneous stopping rule. *Stat. Med.* **2016**, *35*, 5536–5550. [CrossRef]
8. Lookman, T.; Balachandran, P.V.; Xue, D.; Hogden, J.; Theiler, J. Statistical inference and adaptive design for materials discovery. *Curr. Opin. Solid State Mater. Sci.* **2017**, *21*, 121–128. [CrossRef]
9. Shuster, J.J.; Neu, J. A Pocock approach to sequential meta-analysis of clinical trials. *Res. Synth. Methods* **2013**, *4*, 269–279. [CrossRef]
10. Motzer, R.J.; Escudier, B.; Oudard, S.; Hutson, T.E.; Porta, C.; Bracarda, S.; Grünwald, V.; Thompson, J.A.; Figlin, R.A.; Hollaender, N.; et al. Efficacy of everolimus in advanced renal cell carcinoma: A double-blind, randomised, placebo-controlled phase III trial. *Lancet* **2008**, *372*, 449–456. [CrossRef]
11. Goldberg, R.M.; Sargent, D.; Morton, R.F.; Fuchs, C.S.; Ramanathan, R.K.; Williamson, S.K.; Findlay, B.P.; Pitot, H.C.; Alberts, S.R. A Randomized Controlled Trial of Fluorouracil Plus Leucovorin, Irinotecan, and Oxaliplatin Combinations in Patients with Previously Untreated Metastatic Colorectal Cancer. *J. Clin. Oncol.* **2004**, *22*, 23–30. [CrossRef]
12. Bailey, R.C.; Moses, S.; Parker, C.B.; Agot, K.; Maclean, I.; Krieger, J.N.; Williams, C.F.; Campbell, R.T.; Ndinya-Achola, O.J. Male circumcision for HIV prevention in young men in Kisumu, Kenya: A randomized controlled trial. *Lancet* **2007**, *369*, 643–656. [CrossRef]
13. Marcus, R.; Davies, A.; Ando, K.; Klapper, W.; Opat, S.; Owen, C.; Phillips, E.; Sangha, R.; Schlag, R.; Seymour, J.F.; et al. Obinutuzumab for the First-Line Treatment of Follicular Lymphoma. *N. Engl. J. Med.* **2017**, *377*, 1331–1344. [CrossRef]
14. Lui, K.-J. The Performance of the O'Brien-Fleming Multiple Testing Procedure in the Presence of Intraclass Correlation. *Biometrics* **1994**, *50*, 232–236. [CrossRef] [PubMed]
15. Hammouri, H. *Review and Implementation for the O'Brien Fleming Multiple Testing Procedure*; Virginia Commonwealth University: Richmond, VA, USA, 2013. [CrossRef]
16. O'Brien, P.C.; Fleming, T.R. A Multiple Testing Procedure for Clinical Trials. *Biometrics* **1979**, *35*, 549–556. [CrossRef] [PubMed]
17. Thompson, W.R. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* **1933**, *25*, 285–294. [CrossRef]
18. Robbins, H. Some Aspects of the Sequential Design of Experiments. *Bull. Am. Math. Soc.* **1952**, *58*, 527–535. [CrossRef]
19. Zelen, M. Play the Winner Rule and the Controlled Clinical Trial. *J. Am. Stat. Assoc.* **1969**, *64*, 131–146. [CrossRef]
20. Sverdlov, O.; Rosenberger, W.F. On Recent Advances in Optimal Allocation Designs in Clinical Trials. *J. Stat. Theory Pract.* **2013**, *7*, 753–773. [CrossRef]
21. Lehmacher, W.; Wassmer, G. Adaptive Sample Size Calculations in Group Sequential Trials. *Biometrics* **1999**, *55*, 1286–1290. [CrossRef]
22. Liu, Q.; Chi, G.Y.H. On Sample Size and Inference for Two-Stage Adaptive Designs. *Biometrics* **2001**, *57*, 172–177. [CrossRef]
23. Al Garni, H.Z.; Awasthi, A. Al Garni, H.Z.; Awasthi, A. A Monte Carlo Approach Applied to Sensitivity Analysis of Criteria Impacts on Solar PV Site Selection. In *Handbook of Probabilistic Models*, 1st ed.; Samui, P., Bui, D.T., Chakraborty, S., Deo, R.C., Eds.; Butterworth-Heinemann: Oxford, UK, 2020. [CrossRef]
24. Martinez, W.L.; Martinez, A.R. *Computational Statistics Handbook with MATLAB*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2007. [CrossRef]
25. Morris, T.P.; White, I.R.; Crowther, M.J. Using simulation studies to evaluate statistical methods. *Stat. Med.* **2019**, *38*, 2074–2102. [CrossRef] [PubMed]
26. Rosenberger, W.F.; Stallard, N.; Ivanova, A.; Harper, C.N.; Ricks, M.L. Optimal adaptive designs for binary response trials. *Biometrics* **2001**, *57*, 909–913. [CrossRef] [PubMed]
27. Lewis, R.J.; Berry, D.A. Group Sequential Clinical Trials: A Classical Evaluation of Bayesian Decision-Theoretic Designs. *J. Am. Stat. Assoc.* **1994**, *89*, 1528–1534. [CrossRef]
28. Schulz, K.F.; Grimes, A.D. Multiplicity in randomised trials II: Subgroup and interim analyses. *Lancet* **2005**, *365*, 1657–1661. [CrossRef]
29. Pocock, S.J. When (Not) to Stop a Clinical Trial for Benefit. *JAMA* **2005**, *294*, 2228–2230. [CrossRef]

30. Glimm, E.; Maurer, W.; Bretz, F. Hierarchical testing of multiple endpoints in group-sequential trials. *Stat. Med.* **2010**, *29*, 219–228. [CrossRef]
31. Tamhane, A.C.; Mehta, C.R.; Liu, L. Testing a Primary and a Secondary Endpoint in a Group Sequential Design. *Biometrics* **2010**, *66*, 1174–1184. [CrossRef]
32. Tamhane, A.C.; Wu, Y.; Mehta, C.R. Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (I): Unknown correlation between the endpoints. *Stat. Med.* **2012**, *31*, 2027–2040. [CrossRef]
33. Ye, Y.; Li, A.; Liu, L.; Yao, B. A group sequential Holm procedure with multiple primary endpoints. *Stat. Med.* **2013**, *32*, 1112–1124. [CrossRef]
34. Xi, D.; Tamhane, A.C. Allocating recycled significance levels in group sequential procedures for multiple endpoints. *Biom. J.* **2015**, *57*, 90–107. [CrossRef]
35. Lui, K.-J. A Simple Generalization of the O'Brien and Fleming Group Sequential Test Procedure to More Than Two Treatment Groups. *Biometrics* **1993**, *49*, 1216–1219. [CrossRef]