Adaptive Multiple Testing Procedure for Clinical Trials with Urn Allocation

Hanan Hammouri 1,*, Mohammed Ali 1, Marwan Alquran 1, Areen Alquran 2, Ruwa Abdel Muhsen 3 and Belal Alomari 1

1 Department of Mathematics and Statistics, Jordan University of Science and Technology, Irbid 22110, Jordan; myali@just.edu.jo (M.A.); marwan04@just.edu.jo (M.A.); bmalomari207@sci.just.edu.jo (B.A.)
2 Department of Statistics, Yarmouk University, Irbid 21163, Jordan; areen.q@yu.edu.jo
3 Department of Mathematical Sciences, New Mexico State University, Las Cruces, NM 88001, USA; ruwa22@nmsu.edu
* Correspondence: hmhammouri@just.edu.jo

Abstract: This work combines the Urn allocation and O'Brien and Fleming multiple testing procedure to compare two treatments in clinical trials in a novel way. It is shown that this approach overcomes the constraints that previously made it challenging to apply the original adaptive design to clinical trials. The method provides unique flexibility, enabling trials to be stopped early if one treatment shows it is superior without compromising the efficiency of the original multiple testing procedure in terms of type I error rate and power. Experimental data and simulated case examples are used to illustrate the efficacy and robustness of this original approach and its potential for usage in a variety of clinical settings.

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

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

1. Introduction

Over the past decade, the Food and Drug Administration (FDA) has shown concern regarding new challenges faced in conventional experiments. These challenges include difficulty in finding eligible participants, late-stage drug failures, and an increase in the cost of confirmatory trials. These issues have led to a collapse in the drug industry due to economic and ethical factors. In 2004, the FDA showed a decline in medical products submitted for approval during the Critical Path Initiative (CPI). To combat this issue, the FDA proposed that researchers switch from classical designs to adaptive designs [1,2]. Adaptive designs are more flexible and allow for interim analysis tests to be conducted periodically. These tests calculate analysis at the end of each stage and provide the flexibility to end the trial early. Classical designs, however, use all sample sizes within the first stage [1].

The European Medical Agency (EMA) published draft guidelines on various aspects of clinical trials in December 2016 [3]. In January 2017, the Center for Biologics Evaluation and Research (CBER) in the US FDA issued guidance on multiple endpoints in clinical trials; the report was intended for the industry and explained the importance of multiple testing procedures and how alpha should be controlled to prevent type I errors from arising when multiple analyses are conducted [2,4]. In 2018, the FDA published “Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry,” illustrating the importance of using these designs rather than the conventional ones with a primary endpoint [5].

Various testing procedures have been recommended and generalized based on previous recommendations for clinical trials, mainly based on multiple endpoints since these
endpoints significantly influence the design, determination of sample sizes, interim data monitoring, and final analysis [4].

Additionally, several trials showed that a single primary endpoint might only fully represent the overall intervention’s effects after using all the resources [4]. For this reason, many recent clinical trials have relied on multiple testing procedures to finalize their interim analysis. This offers an attractive design feature as they could completely characterize any intervention’s effect without consuming all the resources.

Interim analyses determine whether the trial should be stopped or continued according to the procedures’ critical points [4,6–8]. By terminating the trial as early as possible, analyses can ensure that sources are saved and that candidates are protected from exposure to harmful substances if the trial has a negative outcome [9]. Where early termination for a trial with a positive result means a new product can be used sooner. By making decisions using interim data, these designs can save time, money, and resources while providing valid statistical inferences [10,11].

Several proposals for multiple-testing procedures for endpoints in clinical trials have been proposed in recent years. These procedures were built upon earlier work by Thompson (1933) and O’Brien and Fleming (1979) [12,13], who introduced the concept of the primary key in the interim analysis. O’Brien and Fleming’s work focused on testing the null hypothesis of two equally effective treatments against the restricted alternative that a specified treatment is more effective than the other.

Moreover, Tang, Gnecco, and Geller (1989) also presented an alternative test with a more consistent power function in their work. They proposed an approximate likelihood ratio test for an average mean vector with nonnegative components, which may have been applied in clinical trials or discussed in terms of practical application [14]. As a result, they extended several designs, resulting in new methodologies and guidelines for conducting trials involving multiple endpoints.

Additionally, Lehmacher, Wassmer, and Reitmeir (1991) introduced novel statistical procedures and guidelines to mitigate the issue of increased false-positive findings arising from multiple comparisons with multiple endpoints [15–17]. Afterwards, Tang, Geller, and Pocock (1993) contributed methods for sample size determination, statistical analysis techniques, and strategies to address the correlation between different endpoints.

In subsequent work, Cui, Hung, and Wang (1999) adjusted the sample size in group sequential clinical trials, while Posch and Bauer (1999) developed adaptive designs with conditional error functions. The sample size of sequential group trials was studied by Wassmer and Lehmacher (1999). Their approach involved calculating the sample size at each interim analysis based on the observed data and the planned analysis at the final analysis. Specifically, they used a stopping boundary derived from the O’Brien and Fleming boundary and then adjusted the sample size using a formula incorporating the stopping boundary and the observed variance.

Then, Liu and Chi (2001) proposed an adaptive design method consisting of two stages. The main stage has enough power to reject the null hypothesis, while the extension stage increases the sample size if the true effect size is smaller than anticipated [12,13,18–21]. All these designs and multiple testing procedures can be classified into prospective, concurrent, and retrospective adaptive designs. Collectively, the contributions of these researchers have expanded the methodological repertoire for clinical trials and statistical analysis of multiple endpoints.

Several notable modifications have been made to establish multiple-testing procedures since then. Maurer and Bretz introduced the utilization of correlations in sequential statistics for group-sequential-weighted Bonferroni procedures with multiple endpoints. This advancement enhanced statistical power while controlling the family wise error rate. Building upon their work, Yiyong Fu proposed a Holm-type step-down exact parametric procedure that addressed scenarios with unknown correlations. Additionally, Fu briefly outlined an extension of the semiparametric Seneta-Chen method, which naturally aligns
with the principles of group sequential design. Another relevant contribution came from Zhenming Shun, who investigated the effects of combining sequential group analysis, re-estimation of sample size, and negative stopping (stochastic curtailment) in a single interim analysis focusing on normal data. These modifications offer valuable refinements and expansions to the existing multiple-testing procedures [22–24].

Additionally, Urach and Posch’s contributions to multi-arm sequential designs with the simultaneous stopping rule have brought advancements to the critical boundaries in multi-arm experiments. Their innovative methodologies have improved multi-arm studies’ efficiency, reliability, and decision-making processes, leading to more robust and effective clinical research outcomes [25].

To go back to the basis of this work, O’Brien and Fleming (1979) and Pocock (1977) have served as significant inspirations for the modifications made to sequential group tests [13,26]. Their procedure incorporates type I error control and statistical power, resembling a Chi-square test in the case of a one-stage design. However, it also allows for the early termination of a trial if the treatment demonstrates valuable results achieved through the utilization of critical values.

Researchers in various studies have leveraged the critical values proposed by O’Brien and Fleming. For instance, Hammond et al. [27] have shown that treating COVID-19 patients with nirmatrelvir plus ritonavir early in their illness can significantly decrease the progression to severe disease. Additionally, this treatment has been proven to quickly reduce the SARS-CoV-2 viral load in patients. These critical values highlight the importance of early intervention and provide hope for better outcomes in the fight against COVID-19. Similarly, Goldberg et al. [28] conducted a study to compare three different chemotherapy treatments for metastatic colorectal cancer. They used O’Brien and Fleming’s critical values to monitor the results. The study found that there was no significant difference in overall survival between the three treatments, but two of the treatments, FOLFOX (folinic acid fluorouracil oxaliplatin) and FOLFIRINOX (folinic acid fluorouracil irinotecan oxaliplatin) had higher response rates and longer progression-free survival than the third treatment FOLFIRI (folinic acid fluorouracil irinotecan hydrochloride oxaliplatin). Additionally, Marcus et al. [29] utilized O’Brien and Fleming’s methodology to prematurely conclude gallium trials involving treatment-naive follicular lymphoma patients.

These applications highlight the broad impact and versatility of O’Brien and Fleming’s group sequential testing procedure in the realm of clinical trials, facilitating more efficient decision-making and potentially reducing patient exposure to ineffective or harmful treatments.

The critical values are subsequently used to assess the significance of observed differences between treatment groups and make statistical indications about the treatment effects. Critical values play important roles in the design, analysis, and interpretation of clinical trial results.

Hammouri addressed an important issue in 2013 regarding the stopping bounds of the O’Brien and Fleming procedure, particularly their non-monotonic behavior. To overcome this challenge, Hammouri increased the number of simulations to generate critical values exhibiting monotonic behavior. This adjustment was necessary because critical values should progressively increase with the addition of more interim periods in a clinical trial. Ensuring that critical values follow this pattern makes the control of type I error more robust [30].

In recent years, the statistical method of urn allocation has gained attention for its potential use in clinical trial design. This method involves randomly assigning treatment allocations to patients based on the probability of success for each treatment arm. Compared to traditional fixed allocation designs, urn allocation has been shown to provide greater flexibility and efficiency. However, its use in clinical trials has been limited due to concerns about maintaining the type I error rate and power [31,32].

Randomization is crucial for controlling bias in treatment comparisons. Complete randomization, which is similar to tossing a coin, can minimize or even eliminate bias.
However, in small or moderate-sized trials, complete randomization may lead to a significant imbalance in the number of patients assigned to each group. In fact, according to expert advice [33,34], complete randomization in single-stratum trials with a target sample size of less than 200 is not recommended. There have been various proposed rules for restricted randomization, such as the permuted-block design [35], the biased-coin design and adaptive biased-coin design [36,37], and the urn design [38].

The urn design is a method of randomization that ensures balance in small trials while behaving like complete randomization in larger trials. This means treatment assignments within a sequence generated by the urn design are less predictable than other restricted randomization procedures. This reduces bias risk. In summary, the urn design has unique properties that balance and reduce bias through statistical analysis.

In this study, we present a framework that allows for a more effective comparison of two treatments. The framework integrates the O’Brien and Fleming testing procedure with Urn allocation, resulting in a flexible procedure for effectively detecting differences between treatments. More adaptive features have been incorporated into the procedure, further enhancing its performance. Examples of implementing the new procedure are also provided.

2. Methodologies

In this section, we outline the initial protocols for carrying out clinical trials: the O’Brien and Fleming procedure, as well as the Urn allocation. We then introduce our novel procedure, the Urn Multiple Testing Procedure (UMP), which blends the Urn allocation technique with the O’Brien and Fleming procedure to improve the accuracy of treatment effect assessments while keeping a check on the total type I error rate. Lastly, we delve into the process of evaluating the type I error and statistical power.

2.1. The Original O’Brien and Fleming Procedure

The data are tested and reviewed periodically with $n_1$ and $n_2$ subjects receiving treatment 1 and treatment 2, respectively, with $K$ stages and $N = K(n_1 + n_2)$, such that $N$ is the maximum number of subjects. Then, we use the usual Pearson Chi-square $\chi^2(i)$, a test size, and $P(K, \alpha)$, which is the O’Brien and Fleming’s critical value in Table 1.

<table>
<thead>
<tr>
<th>K</th>
<th>1.0000</th>
<th>2.0000</th>
<th>3.0000</th>
<th>4.0000</th>
<th>5.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.462</td>
<td>0.656</td>
<td>0.75</td>
<td>0.785</td>
<td>0.819</td>
</tr>
<tr>
<td></td>
<td>0.4547</td>
<td>0.6546</td>
<td>0.7439</td>
<td>0.8013</td>
<td>0.8431</td>
</tr>
<tr>
<td>0.1</td>
<td>2.67</td>
<td>2.859</td>
<td>2.907</td>
<td>2.979</td>
<td>3.087</td>
</tr>
<tr>
<td></td>
<td>2.7042</td>
<td>2.8195</td>
<td>2.9247</td>
<td>3.0047</td>
<td>3.0650</td>
</tr>
<tr>
<td>0.09</td>
<td>2.866</td>
<td>3.031</td>
<td>3.073</td>
<td>3.147</td>
<td>3.283</td>
</tr>
<tr>
<td></td>
<td>2.8730</td>
<td>2.9817</td>
<td>3.0877</td>
<td>3.1668</td>
<td>3.2225</td>
</tr>
<tr>
<td>0.08</td>
<td>3.077</td>
<td>3.197</td>
<td>3.24</td>
<td>3.338</td>
<td>3.467</td>
</tr>
<tr>
<td></td>
<td>3.0633</td>
<td>3.1646</td>
<td>3.2700</td>
<td>3.3498</td>
<td>3.4114</td>
</tr>
<tr>
<td>0.07</td>
<td>3.294</td>
<td>3.363</td>
<td>3.437</td>
<td>3.546</td>
<td>3.663</td>
</tr>
<tr>
<td></td>
<td>3.2814</td>
<td>3.3754</td>
<td>3.4799</td>
<td>3.5594</td>
<td>3.6220</td>
</tr>
<tr>
<td>0.06</td>
<td>3.576</td>
<td>3.652</td>
<td>3.683</td>
<td>3.853</td>
<td>3.889</td>
</tr>
<tr>
<td></td>
<td>3.5348</td>
<td>3.6207</td>
<td>3.7251</td>
<td>3.8034</td>
<td>3.8669</td>
</tr>
<tr>
<td>0.05</td>
<td>3.869</td>
<td>3.928</td>
<td>3.94</td>
<td>4.17</td>
<td>4.149</td>
</tr>
<tr>
<td></td>
<td>3.8399</td>
<td>3.9152</td>
<td>4.0191</td>
<td>4.0961</td>
<td>4.1602</td>
</tr>
<tr>
<td>0.04</td>
<td>4.289</td>
<td>4.231</td>
<td>4.264</td>
<td>4.477</td>
<td>4.584</td>
</tr>
<tr>
<td></td>
<td>4.2177</td>
<td>4.2809</td>
<td>4.3836</td>
<td>4.4599</td>
<td>4.5243</td>
</tr>
</tbody>
</table>
In any $i$ stage, if $\chi^2_{(i)} \geq P(K, \alpha)$, the null hypothesis is rejected, and the experiment is terminated. Otherwise, if the critical value does not exceed $P(K, \alpha)$, the study is terminated by concluding that the hypothesis of no difference cannot be rejected at the ($\alpha$) significance level.

For further details about computing and estimating the original and corrected critical points, see [30,39].

2.2. Urn Allocation

Urn allocation was given in Zelen [40], Wei and Durham [41], and Ivanovna [42] as $\delta$ is simply the fraction of patients on treatment 1. Where $
abla \delta(p_1, p_2) = \frac{1 - p_2}{2 - p_1 - p_2} = \frac{q_2}{q_1 + q_2}$ where $p_i$ represents the success probability of a patient on treatment $i$. Additionally, $q_i = 1 - p_i$ represents the failure probability for $i = 1, 2$. In a clinical trial comparing two treatments (treatment 1 and treatment 2) with binary responses (success and failure), patients are enrolled sequentially, and their responses to the treatments are observed immediately. This study design, famously known as the play-the-winner (PW) rule, was introduced by Zelen in 1969. According to this rule, if a patient experiences success with a specific treatment, the subsequent patient will receive the same treatment, while if a patient’s treatment is a failure, the next patient will be assigned to the alternate treatment. Then

$$\frac{N_{n,1}}{n} \overset{p}{\rightarrow} \frac{q_1}{q_1 + q_2}$$

and

$$\sqrt{n}\left(\frac{N_{n,1}}{n} - \frac{q_2}{q_1 + q_2}\right) \overset{D}{\rightarrow} N\left(0, \sigma_{PW}^2\right),$$

where $N_{n,k}$ be the number of subjects assigned to treatment $k$ in the first $n$ assignments and $N_n = (N_{n,1}, \ldots, N_{n,K})$. The statistical behavior of proportions $N_{n,K}/n, k = 1, \ldots, K$ is of interest. Moreover, $\sigma_{PW}^2 = \frac{n q_2 (p_1 + p_2)}{(q_1 + q_2)^3}$, since $\frac{q_2}{(q_1 + q_2)} > \frac{1}{2}$ if $p_1 > p_2$. If treatment 1 is better, the PW rule favors treatment 1 [43].

Another researcher reached the same result with more details [44], where the probability that the superior treatment is allocated to each patient would be more than 0.5 under the use of an adaptive design. Although its demonstration may be difficult in general, they showed it for the deterministic play-the-winner design (Zelen [40]). Under this design, the first patient is allocated to either treatment by a simple randomization. Then the same treatment is applied after a success, and the other treatment is used after a failure. Assuming clinical equipoise (Freedman [45]), the principle of interchangeability is satisfied when the initial patient is enrolled.
Suppose the deterministic play-the-winner rule is followed. Let \( \Delta = p_1 - p_2 \) and \( K = p_1 + p_2 \). To avoid triviality, assume that \( K \neq 2 \). Then \( p_1 = 1/2 \) and for any integer \( n \geq 0 \), by mathematical induction,

\[
p_{n+1} = \frac{1}{2} + \frac{1}{2} \Delta + \frac{1}{2} \Delta(k-1) + \cdots + \frac{1}{2} \Delta (k-1)^{n-1} = \frac{1}{2} + \frac{1}{2} \frac{1 - (k-1)^{n}}{2-k}
\]

This sequence of probabilities of allocations on treatment 1 has the following properties.

1. \( p_{n+1} > \frac{1}{2} \) for any integer \( n \geq 0 \), if \( p_1 > p_2 \), \( p_{n+1} < \frac{1}{2} \) if \( p_1 < p_2 \), and \( p_{n+1} = 0.5 \) if \( p_1 = p_2 \). Moreover, when \( K > 1 \), \( p_{n+1} \) is increasing in \( n \) when \( p_1 > p_2 \), and it is decreasing when \( p_1 < p_2 \).

2. \( P(p_{n+1} > \frac{1}{2}) > P(p_1 > p_2) \).

3. \( p = \lim_{n \to \infty} p_{n+1} = \frac{(1-p_2)}{(1-p_1)+(1-p_2)} = \frac{q_2}{q_1+q_2} \) exists since \( p_n \) is a Cauchy sequence.

4. \( p > \frac{1}{2}, \text{ if } p_1 > p_2 \).

5. \( P(p > \frac{1}{2}) = P(p_1 > p_2) \).

6. \( \lim_{p_1 \to 1} p = 1 \), \( \lim_{\Delta \to 0} p = \frac{1}{2} \), \( \lim_{p_2 \to 1} p = 0 \).

7. If \( K = 1 \), then \( p_{n+1} = p = p_1 \).

Properties 1, 2, 4, and 5 say that the probability of using the superior treatment is more than 50%. Property 3 indicates that \( p \) is simply the asymptotic fraction of patients on treatment 1 (Here \( p \) is the same \( \delta \)). This outcome aligns with the findings of Zelen [40] and Wei and Durham [41]. Property 6 shows that if \( p_1 \) (or \( p_2 \)) is sufficiently large, treatment 1 (or 2) will eventually be identified as the superior treatment. On the other hand, if there is no difference between the two treatments, we eventually randomize between them. Property 7 assumes that \( p_2 = 1 - p_1 \). Suppose that \( p_1 > 0.5 \). If we call it a “success” if treatment 1 is allocated and a “failure” if 2 is used, we essentially have a binomial experiment with \( p_1 \) as the probability of success.

2.3. The Proposed Method: Urn Multiple Testing Procedure (UMP)

Our framework combines Urn allocation with the O’Brien and Fleming procedure, enabling researchers to monitor and analyze data collected at each interim analysis stage. This empowers them to make informed decisions based on predefined stopping rules and efficacy boundaries. Our procedure integrates the strengths of Urn allocation and the O’Brien and Fleming procedure, resulting in a comprehensive and sophisticated methodology that enhances the reliability and efficiency of studies. By leveraging these methods, our procedure leads to more meaningful and accurate results, ultimately providing a valuable contribution to the field.

The New Methodology

The new procedure involves periodically reviewing and testing collected data in \( K \) stages, with a designated sample size, denoted as \( n_i \) where \( i = 1, \ldots, K \). The sample sizes will be determined using equal allocation.

In order to carry out the Urn allocation method, we split the sample size \( n_i \) into two portions, \( n_{i1} \) and \( n_{i2} \), which will be assigned to treatment 1 and treatment 2, respectively. This allocation guarantees that the sum of \( n_{i1} \) and \( n_{i2} \) equals the overall sample size, \( n_i \).

The equal allocation method will be used to determine the subsample sizes in the first stage. This entails dividing the total sample size \( n_1 \) evenly into two subgroups. The Urn allocation method will be employed in subsequent stages to determine the subsample sizes. This procedure takes into account data from previous stages and adjusts the allocation based on observed treatment outcomes. By using this method, more participants are allocated to the treatment with a greater chance of success, resulting in increased study efficiency.
The O’Brien and Fleming procedure will be used for the other steps to determine whether to continue or stop testing. This process involves comparing the observed test statistic, which is usually the Pearson Chi-square $\chi^2$, with the corrected critical values provided in Table 1. These values set the thresholds for statistical significance at each study stage.

If the observed test statistic exceeds the critical value, indicating a significant treatment effect, we may terminate the test early and conclude that the treatment is effective. Conversely, if the observed test statistic falls below the critical value, suggesting no significant treatment effect, we may choose to continue the test and collect additional data to gather more evidence.

1. To begin, consider the overall sample size ($N$) and the number of stages involved ($K$). It is important to note that $N$ should be an even number, as equal allocation may be used if Urn allocation fails, and this requires even sample sizes across stages.

2. The process proceeds first by finding each stage sample size. By calculating $n_i = \text{round} \left( \frac{N}{K} \right)$ if $n_1$ is even. Otherwise, $n_1 = \text{round} \left( \frac{N}{K} \right) + 1$, for $i = 1, \ldots, K - 1$ and $n_K = N - \sum_{i=1}^{K-1} n_i$.

3. For $i = 1$, the sample size for the first stage $n_1$ and equal allocation is used to compute the subsamples, where $n_{11} = n_{12} = \frac{N}{2}$.

4. For $i = 2, \ldots, K$, $n_i$ is calculated as the subsample given by $n_{i1} = \text{round}(y_i \cdot n_i)$ with $y_i = \frac{q_{i-1,2}}{q_{i-1,1}+q_{i-1,2}}$, where $q_{i-1,j} = 1 - p_{i-1,j}$ and $p_{1-1,j}$ for $j = 1, 2$ are success rates in the cumulative previous stages for treatment 1 and treatment 2, respectively, and $n_{i2} = (n_i - n_{i1})$. In case where $n_{11}$ or $n_{12}$ equals zero, equal allocation will be used.

5. At that time, in each stage $i$, starting from the first one, subjects are randomized, and their measurements are observed. Each subsample will be added up to the previous subsamples for the same treatment. $\frac{1}{K} \chi^2_{(i)}$ is calculated and compared to $P(K, \alpha)$.
   i. If $\frac{1}{K} \chi^2_{(i)} \geq P(K, \alpha)$, then the study ends, and the null hypothesis is rejected. Otherwise, if $\frac{1}{K} \chi^2_{(i)} < P(K, \alpha)$, and $i < K$, the procedure precedes to the next stage.
   ii. If $\frac{1}{K} \chi^2_{(i)} < P(K, \alpha)$ and $i = K$, the study is terminated, and the null hypothesis fails to be rejected.

The methodology for the new procedure is graphically illustrated in the following flow chart, Figure 1.

2.4. Type I Error and Power

The purpose of this section is to prove the accuracy of the procedure by examining the new type I error and power. Monte Carlo simulation will be used to assess the effectiveness of a procedure. This method helps us to examine the power and the error rate for conducting the new procedure, UMP, and compare it to the original procedure. With Monte Carlo simulations, we can demonstrate how well the UMP performs in detecting significant effects and ensure that it maintains the same type I error rate, which upholds statistical standards.

2.4.1. Type I Error Procedure

The simulation using SAS (Statistical Analysis System, version 9.4 (SAS Institute Inc., Cary, NC, USA)) code was used to calculate type I errors to evaluate the new procedure. The simulations take in a range of success probabilities ($p = 0.1, 0.2, 0.3, 0.4, 0.5$), test sizes ($\alpha = 0.99$ and $0.95$), and critical values ($P(K, \alpha)$) for $K$ values ranging from 1 to 5.

By conducting these simulations, the performance of the UMP under different scenarios can be evaluated. The varying success probabilities will allow an exploration of the robustness and adaptability to a wide range of conditions, furthermore, considering different test sizes and critical values.
2.4. Type I Error and Power

The purpose of this section is to prove the accuracy of the procedure by examining the new type I error and power. Monte Carlo simulation will be used to assess the effectiveness of a procedure. This method helps us to examine the power and the error rate for conducting the new procedure, UMP, and compare it to the original procedure. With Monte Carlo simulations, we can demonstrate how well the UMP performs in detecting significant effects and ensure that it maintains the same type I error rate, which upholds statistical standards.

2.4.1. Type I Error Procedure

In order to determine if the new procedure is effective at rejecting the null hypothesis or detecting a significant difference between groups, we will create multiple subsamples from a single binomial distribution with \( p = 0.1, 0.2, 0.3, 0.4, \) and \( 0.5 \) with sample sizes \( N = 120, 300, \) and \( 720 \) for \( \alpha = 0.01 \) and with \( N = 20, 360, \) and \( 720 \) for \( \alpha = 0.05 \). Equal success rates were used (ensuring no difference can be found and the \( H_0 \) is true). This process will be repeated 500,000 times, and we will calculate the percentage of times the \( H_0 \) is rejected to determine the type I error rate.

We created Figure 2 to visually represent the process of determining the type I error for the UMP.
The simulation using SAS (Statistical Analysis System, version 9.4 (SAS Institute Inc., Cary, NC, USA)) code was used to calculate type I errors to evaluate the new procedure. The simulations take in a range of success probabilities ($p = 0.1, 0.2, 0.3, 0.4, 0.5$), test sizes ($\alpha = 0.99$ and $0.95$), and critical values ($P(K, \alpha)$) for $K$ values ranging from 1 to 5. By conducting these simulations, the performance of the UMP under different scenarios can be evaluated. The varying success probabilities will allow an exploration of the robustness and adaptability to a wide range of conditions, furthermore, considering different test sizes and critical values.

In order to determine if the new procedure is effective at rejecting the null hypothesis or detecting a significant difference between groups, we will create multiple subsamples from a single binomial distribution with $p = 0.1, 0.2, 0.3, 0.4, 0.5$ with sample sizes $N = 120, 300, 720$ for $\alpha = 0.01$ and with $N = 20, 360, 720$ for $\alpha = 0.05$. Equal success rates were used (ensuring no difference can be found and the $H_0$ is true). This process will be repeated 500,000 times, and we will calculate the percentage of times the $H_a$ is rejected to determine the type I error rate.

We created Figure 2 to visually represent the process of determining the type I error for the UMP.

2.4.2. Power Procedure

The SAS program was employed to determine the power values for a specific analysis. In the initial step, a probability value, $P_1$, was selected as 0.1, along with success probability values, $P_2$, of 0.15, 0.2, 0.25, and 0.3. The program executed 500,000 iterations with significance levels $\alpha = 0.01$ and $\alpha = 0.05$. The analysis incorporated the corrected O’Brien and Fleming critical values, $P(K, \alpha)$. Using the conventional Chi-square test, sample sizes were determined to achieve power values around 0.8.

For $\alpha = 0.05$, the sample sizes were 1380, 396, 200, and 120, while for $\alpha = 0.01$, the sizes were 2040, 600, 300, and 180, to ensure a significant difference. Two subsamples were generated for each case of $K = 1, \ldots, 5$ from different binomial distributions with different means. The new procedure was utilized to test whether the null hypothesis ($H_0$) indicating that there is no difference between the two groups, could be rejected in favor of the alternative hypothesis ($H_a$). The entire process was repeated 500,000 times to calculate the proportion of rejections of $H_0$, representing the power rate, given that $H_a$ was guaranteed to be true. The graphical representation of the power computation process can be found in Figure 2.

3. Results

This section will present the results of our analysis of the type I error and power levels for the proposed procedure. Our findings show that the type I error rate was acceptable, indicating that we did not reject the null hypothesis when it was true. Additionally, the
power level of the test was high, suggesting that we were able to accept the alternative hypothesis when it was true correctly. These results demonstrate the validity and reliability of our proposed procedure and provide confidence in the accuracy of our findings.

3.1. Type I Error Results

After conducting an extensive analysis of various sample sizes, it was observed that the results remained consistent irrespective of the sample size. Furthermore, when comparing these results to the conventional Chi-square test, it was found that the modified procedure reduced type I error values. This reduction became more pronounced as the value of $K$ increased.

The rationale behind this trend lies in the fact that as $K$ increases, the Chi-square statistic is multiplied by a factor that is less than one. Consequently, the resulting value is compared with an increasingly larger critical value. As a result, it becomes more difficult to reject the null hypothesis $H_0$. In other words, the larger the value of $K$, the stronger the evidence required to reject the null hypothesis.

The results for sample sizes 360 and 300 are presented in Table 2, with significance levels of 0.05 and 0.01, respectively. In the first case, where the sample size was 360, and the coefficient of determination was 0.05, we observed that the type I error ranges from 0.04907 to 0.05080 when $K = 2$. For $K = 5$, the type I error values range from 0.048838 to 0.05047, which are very close to the significance level of 0.05. This improvement in error control over the usual Chi-square procedure is considered acceptable.

The type I error also showed similar patterns when considering the same significance level (alpha) and varying sample sizes of 120 and 720. For the sample size of 120, the type I error ranges from 0.042836 to 0.051654. For the sample size 720, the type I error ranges from 0.049216 to 0.05063.

When using a significance level of $\alpha = 0.01$, the values range between 0.008882 and 0.09922 in the second stage and range between 0.008866 and 0.00976 in the last stage. Importantly, these values remain below 0.01, which is considered satisfactory as the errors do not exceed this threshold.

Furthermore, we calculated the type I error values for sample sizes 120 and 720 using a significance level of $\alpha = 0.01$. The results indicated that the new procedure performs appropriately in terms of type I errors. Specifically, for a sample size of 120, all type I error values fall within the range of 0.008394 to 0.010064. Similarly, for a sample size of 720, the type I error values range from 0.009494 to 0.010064.

Tables 2–4 provide a comprehensive overview of the remaining type I error values for different sample sizes and significance levels, allowing for a thorough examination of the
procedure’s performance in various scenarios. Figure 3a,b shows the type I error values from the 500,000 simulations.

Table 3. Type I error of UMP with $\alpha = 0.01, 0.05$, and sample size 120.

<table>
<thead>
<tr>
<th>K</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.050698</td>
<td>0.050698</td>
</tr>
<tr>
<td></td>
<td>0.0088354</td>
<td>0.0088354</td>
</tr>
<tr>
<td>0.2</td>
<td>0.05349</td>
<td>0.05349</td>
</tr>
<tr>
<td></td>
<td>0.009372</td>
<td>0.009372</td>
</tr>
<tr>
<td>0.3</td>
<td>0.052446</td>
<td>0.052446</td>
</tr>
<tr>
<td></td>
<td>0.009372</td>
<td>0.009372</td>
</tr>
<tr>
<td>0.4</td>
<td>0.05065</td>
<td>0.05065</td>
</tr>
<tr>
<td></td>
<td>0.009372</td>
<td>0.009372</td>
</tr>
<tr>
<td>0.5</td>
<td>0.05446</td>
<td>0.05446</td>
</tr>
<tr>
<td></td>
<td>0.009372</td>
<td>0.009372</td>
</tr>
</tbody>
</table>

Table 4. Type I error of UMP with $\alpha = 0.01, 0.05$, and sample size 720.

<table>
<thead>
<tr>
<th>K</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.05223</td>
<td>0.05223</td>
</tr>
<tr>
<td></td>
<td>0.011272</td>
<td>0.011272</td>
</tr>
<tr>
<td>0.2</td>
<td>0.051084</td>
<td>0.051084</td>
</tr>
<tr>
<td></td>
<td>0.010432</td>
<td>0.010432</td>
</tr>
<tr>
<td>0.3</td>
<td>0.05021</td>
<td>0.05021</td>
</tr>
<tr>
<td></td>
<td>0.010082</td>
<td>0.010082</td>
</tr>
<tr>
<td>0.4</td>
<td>0.051684</td>
<td>0.051684</td>
</tr>
<tr>
<td></td>
<td>0.010222</td>
<td>0.010222</td>
</tr>
<tr>
<td>0.5</td>
<td>0.048062</td>
<td>0.048062</td>
</tr>
<tr>
<td></td>
<td>0.010162</td>
<td>0.010162</td>
</tr>
</tbody>
</table>

Figure 3. Type I error of UMP with (a) $\alpha = 0.01$ and sample size 300, (b) $\alpha = 0.05$ and sample size 360.
3.2. Power Results

As a result, it was noticed that the initial values for the power value with \( K = 1 \) and \( \alpha = 0.05 \) were between 0.80346 and 0.81579. Moreover, the values showed a decreased behavior when the values of \( K \) were increased, since the power values with \( \alpha = 0.05 \) when \( K = 5 \) were between 0.78067 and 0.79628, with marginal errors less than 0.01933 between the power values with \( K = 5 \) and 0.8.

Then, power values with an \( \alpha \) of 0.01 and \( K \) value of 1 range from 0.80087 to 0.81310. Upon further analysis, we found again that the power values decreased as \( K \) values increased. Specifically, when \( K \) was equal to 5, the power values ranged from 0.78204 to 0.8030, with marginal errors less than 0.01796 between the power values with \( K = 5 \) and 0.8. In both cases, the marginal error is very small and can be overlooked compared to the benefits gained from using this method.

The remaining values and power values behaviors have been presented in Table 5 and visually represented in Figure 4a,b.

Table 5. Power estimates for different cases for implementing Urn allocation and equal weights to the O’Brien Fleming procedure with \( \alpha = 0.05 \) and \( \alpha = 0.01 \).

<table>
<thead>
<tr>
<th>( K )</th>
<th>Total Sample Sizes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>1380</td>
<td>0.80346</td>
<td>0.80034</td>
<td>0.79668</td>
<td>0.79267</td>
<td>0.79342</td>
</tr>
<tr>
<td></td>
<td>2040</td>
<td>0.80087</td>
<td>0.80028</td>
<td>0.79734</td>
<td>0.79716</td>
<td>0.79413</td>
</tr>
<tr>
<td>2</td>
<td>396</td>
<td>0.80646</td>
<td>0.79934</td>
<td>0.79534</td>
<td>0.79365</td>
<td>0.79082</td>
</tr>
<tr>
<td></td>
<td>600</td>
<td>0.81216</td>
<td>0.80923</td>
<td>0.8059</td>
<td>0.80373</td>
<td>0.80300</td>
</tr>
<tr>
<td>0.25</td>
<td>200</td>
<td>0.81579</td>
<td>0.80616</td>
<td>0.80055</td>
<td>0.79946</td>
<td>0.79628</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.81310</td>
<td>0.81184</td>
<td>0.80655</td>
<td>0.80435</td>
<td>0.80177</td>
</tr>
<tr>
<td>0.3</td>
<td>120</td>
<td>0.81287</td>
<td>0.79706</td>
<td>0.79117</td>
<td>0.78209</td>
<td>0.78067</td>
</tr>
<tr>
<td></td>
<td>180</td>
<td>0.80169</td>
<td>0.79713</td>
<td>0.7932</td>
<td>0.79021</td>
<td>0.78204</td>
</tr>
</tbody>
</table>

Figure 4. Power values of UMP with (a) \( \alpha = 0.05 \), (b) \( \alpha = 0.01 \), \( P_1 = 0.1 \), and \( P_2 = 0.15, 0.2, 0.25, 0.3 \).

3.3. Calculating Rejection Rates for Each Stage

In this section, our focus was to determine at which stage the rejection occurs by calculating the rejection rates and identifying the sample size required to conclude the rejection of the null hypothesis (\( H_0 \)).

3.3.1. Calculating Rejection Rates for Each Stage When the Difference Is Present

With a standard power of 0.8 and considering different probabilities of success (0.1 and 0.15) with an \( \alpha = 0.05 \) threshold, a total sample size of 1380 was determined. Using
500,000 iterations at each stage ($i$), the results were summarized in Table 6 and illustrated in Figure 5.

Table 6. Sample sizes and percentages of accepting $H_0$ occurring at stage $k$ with 500,000 iterations.

<table>
<thead>
<tr>
<th>$k$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Stage</td>
<td>690 (25.7%)</td>
<td>460 (3.7%)</td>
<td>346 (0.4%)</td>
<td>276 (0%)</td>
</tr>
<tr>
<td>Second Stage</td>
<td>1380 (74.3%)</td>
<td>920 (50.9%)</td>
<td>692 (23%)</td>
<td>552 (8.6%)</td>
</tr>
<tr>
<td>Third Stage</td>
<td>1380 (45.4%)</td>
<td>1036 (45%)</td>
<td>828 (32.2%)</td>
<td></td>
</tr>
<tr>
<td>Fourth Stage</td>
<td>1380 (31.6%)</td>
<td>1104 (35%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fifth Stage</td>
<td>1380 (24.2%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. Percent of rejections of $H_0$ occurring at stage $i$ with 500,000 iterations.

3.3.2. Calculating Rejection Rates for Each Stage When the Difference Is Not Presented

The rejections were calculated with $\alpha = 0.01$, and the sample size was 360. Based on 500,000 iterations, Table 7 and Figure 6 illustrate the required sample size and the number of rejections at stage $K$.
Table 7. Sample sizes and percentages of rejecting $H_0$ occurring at stage $i$ with 500,000 iterations.

<table>
<thead>
<tr>
<th>$K$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Stage</td>
<td>180 (1.5%)</td>
<td>120 (0%)</td>
<td>90 (0%)</td>
<td>72 (0%)</td>
</tr>
<tr>
<td>Second Stage</td>
<td>360 (98.5%)</td>
<td>240 (11.1%)</td>
<td>180 (6.2%)</td>
<td>144 (0%)</td>
</tr>
<tr>
<td>Third Stage</td>
<td>360 (88.9%)</td>
<td>270 (33.1%)</td>
<td>216 (4.5%)</td>
<td></td>
</tr>
<tr>
<td>Fourth Stage</td>
<td>360 (60.7%)</td>
<td>288 (28.7%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fifth Stage</td>
<td>360 (66.8%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. Percent of rejections of $H_0$ occurring at stage $i$ with 500,000 iterations.

Moreover, it needs to be noted that these percentages are out of the 5% rejection 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.

4. Examples

4.1. Example 1: Computational Example

We conducted a trial on simulated data of 1200 patients divided into two groups (600 each). Group 1 data was simulated using a binomial distribution with a success rate of 0.25, and group 2 data was simulated using a binomial distribution with a success rate of 0.55.

The UMP was used to find the significant association between treatment in use and success for $K = 1$ to 5 and alpha of 0.05. We analyzed the data and presented case five as part of our findings to illustrate calculations.

For case five, the UMP was used with sample sizes equal to $n_1 = 240$ and a critical value of 4.1602 (from Table 1).

For the first stage, with $n_1 = 240$, and the subsamples distributed equally with $n_{11} = n_{12} = 120$. Then, the Chi-Square statistic was 8.6, and after dividing the Chi-Square value by one-fifth, the value was 1.72, which is not greater than the critical value, so we failed to reject the $H_0$ hypotheses.

For the second stage, with $n_2 = 240$, the cumulative subsamples were distributed by using Urn allocation as follows:

Since $p_{1,1} = \frac{34}{120} = 0.28$, and $p_{1,2} = \frac{56}{120} = 0.47$. Then, we got $y_2 = \frac{1-p_{1,2}}{p_{1,1} + p_{1,2}} = \frac{0.53}{1.25} = 0.424$, with $n_{21} = y_2 * n_2 = Round(101.76) = 102$, then, the cumulative $= n_{11} + n_{21} = 120 + 102 = 222$, and the cumulative $= n_{12} + n_{22} = 258$. The successes were counted, and the Chi-square statistic was calculated.
The Chi-Square statistic was 23.35, and after dividing the Chi-Square value by two-fifths, the value was 9.34, which is greater than the critical value, so we rejected the $H_0$ hypotheses.

In total, the experiment was terminated after two stages with 480 participants out of 1200 patients due to significant differences between the two groups.

The complete information regarding each $K$ is presented in Table 8.

### Table 8. The results of the UMP for the computational example.

<table>
<thead>
<tr>
<th>$i$</th>
<th>Critical Values</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>Total Sample Size</th>
<th>Chi-Square</th>
<th>$\frac{i}{K}X_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.84</td>
<td>600</td>
<td>600</td>
<td>153</td>
<td>296</td>
<td>1200</td>
<td>72.77</td>
<td>72.77</td>
</tr>
<tr>
<td></td>
<td>Result of case $K = 1$ ($n = 1200$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.92</td>
<td>300</td>
<td>300</td>
<td>83</td>
<td>143</td>
<td>600</td>
<td>25.56</td>
<td>12.78</td>
</tr>
<tr>
<td></td>
<td>Result of case $K = 2$, ($n_1 = n_2 = 600$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.94</td>
<td>200</td>
<td>200</td>
<td>54</td>
<td>98</td>
<td>400</td>
<td>20.77</td>
<td>6.92</td>
</tr>
<tr>
<td></td>
<td>Result of case $K = 3$ ($n_1 = n_2 = n_3 = 400$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.10</td>
<td>150</td>
<td>150</td>
<td>41</td>
<td>69</td>
<td>300</td>
<td>11.25</td>
<td>2.8125</td>
</tr>
<tr>
<td></td>
<td>Result of case $K = 4$ ($n_1 = n_2 = n_3 = n_4 = 300$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4.10</td>
<td>278</td>
<td>322</td>
<td>75</td>
<td>157</td>
<td>600</td>
<td>29.84</td>
<td>14.92</td>
</tr>
<tr>
<td></td>
<td>Result of case $K = 5$ ($n_1 = n_2 = n_3 = n_4 = n_5 = 240$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.16</td>
<td>120</td>
<td>120</td>
<td>34</td>
<td>56</td>
<td>240</td>
<td>8.6</td>
<td>1.72</td>
</tr>
<tr>
<td>2</td>
<td>4.16</td>
<td>222</td>
<td>258</td>
<td>59</td>
<td>124</td>
<td>480</td>
<td>23.35</td>
<td>9.34</td>
</tr>
</tbody>
</table>

In this example, the trial can be terminated using only 400 of 1200 patients, as shown in Figure 7.

![Figure 7. Sizes of samples necessary to reach the rejection of $H_0$ for the computational example.](image-url)
4.2. Example 2: Experimental Data Example

For this example, an experimental study was conducted by distributing a brief survey among university students in April 2023 at Jordan University of Science and Technology. The purpose of this study was solely illustrative, and no inferential conclusions were intended to be drawn from the data. The study aimed to demonstrate our methodology, which was designed to minimize the number of participants required for the experiment. The primary focus was to investigate the association between children’s smoking habits and their parents’ smoking history.

The survey data included 240 participants, with 120 smokers and 120 non-smokers in each group. Each participant was asked about their smoking status, and the smoking histories of their parents were recorded for analysis. Once 120 participants were reached for each group, data collection stopped. A total of 386 participants were sampled to reach this count.

Instead of utilizing the entire sample (i.e., \( K = 1 \)), the UMP method was employed to showcase that the necessary results could be obtained with fewer than 240 participants. The data was used for \( K = 2, 3, 4, \) and \( 5 \) and alpha of 0.05. In each scenario, each subsample was further divided into \( K \) stages, and UMP was applied.

To demonstrate the procedure, we included details of \( K = 4 \):

A conclusion of significance was reached after two stages of the trial, with 120 out of 240 participants.

In detail:
- For the first stage with \( K = 4 \), we have \( n_1 = 60 \), and the subsample was divided equally, where \( n_{11} = n_{12} = 30 \), so the Chi-square statistic was 8.5, and by multiplying the Chi-square by one-fourth, the value was 2.13, which is not greater than the critical value of 4.0961, so we failed to reject \( H_0 \) hypothesis.
- For the second stage, we needed to recalculate \( n_{21} \) and \( n_{22} \) by using the Urn allocation. We have the subsamples equal to \( n_{21} = 46 \) and \( n_{22} = 14 \), with cumulative subsamples equal to 76 and 44 for group1 and group2 respectively. Moreover, two-fourths of the Chi-square statistic was 15.81, which is greater than the critical value. Thus, we reject the \( H_0 \) hypothesis.

In this case, we used only 120 patients to end the experiment and get a significant difference between the two treatments.

For more details regarding each \( K \), see Table 9.

<table>
<thead>
<tr>
<th>( i )</th>
<th>Critical Values</th>
<th>( n_1 )</th>
<th>( n_2 )</th>
<th>( \chi_1 )</th>
<th>( \chi_2 )</th>
<th>Total Sample Size</th>
<th>Chi-Square</th>
<th>( \frac{i - 1}{K} \chi_1^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.84</td>
<td>120</td>
<td>120</td>
<td>105</td>
<td>37</td>
<td>240</td>
<td>79.75</td>
<td>79.75</td>
</tr>
<tr>
<td></td>
<td>Result of case 1 ((K = 1)) ( n = 240 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.92</td>
<td>60</td>
<td>60</td>
<td>51</td>
<td>17</td>
<td>120</td>
<td>39.23</td>
<td>19.62</td>
</tr>
<tr>
<td></td>
<td>Result of case 2 ((K = 2)) ( (n_1 = n_2 = 120) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.94</td>
<td>40</td>
<td>40</td>
<td>36</td>
<td>13</td>
<td>80</td>
<td>27.86</td>
<td>9.29</td>
</tr>
<tr>
<td></td>
<td>Result of case 3 ((K = 3)) ( (n_1 = n_2 = n_3 = 80) )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.02</td>
<td>30</td>
<td>30</td>
<td>22</td>
<td>7</td>
<td>60</td>
<td>15.017</td>
<td>3.75</td>
</tr>
<tr>
<td>2</td>
<td>4.02</td>
<td>76</td>
<td>44</td>
<td>63</td>
<td>14</td>
<td>120</td>
<td>31.62</td>
<td>15.81</td>
</tr>
</tbody>
</table>

Table 9. The results of the UMP for the experimental data example.
Table 9. Cont.

<table>
<thead>
<tr>
<th>i</th>
<th>Critical Values</th>
<th>n₁</th>
<th>n₂</th>
<th>χ₁</th>
<th>χ₂</th>
<th>Total Sample Size</th>
<th>Chi-Square</th>
<th>$\frac{i}{K}\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.16</td>
<td>24</td>
<td>24</td>
<td>17</td>
<td>7</td>
<td>48</td>
<td>8.33</td>
<td>1.67</td>
</tr>
<tr>
<td>2</td>
<td>4.16</td>
<td>58</td>
<td>38</td>
<td>49</td>
<td>12</td>
<td>96</td>
<td>27.737</td>
<td>11.09</td>
</tr>
</tbody>
</table>

Result of case 5 (K = 5) 

$n₁ = n₂ = n₃ = n₄ = n₅ = 48$

Similarly, the trial can be terminated in the second example using only 80 of 240 patients, as shown in Figure 8.

5. Discussion

Clinical trials aim to enhance efficiency by employing sequential procedures that incorporate multiple primary critical points. By considering multiple critical endpoints, trials can comprehensively characterize treatment effectiveness. Several forms of these procedures have been proposed in order to optimize the analysis.

O’Brien and Fleming proposed one of the earliest multiple-testing procedures in 1979. However, it has been noted by Hammouri [30] that the original critical values of this procedure required modification to ensure monotonic behavior after 1,000,000 iterations. This study specifically combined the O’Brien and Fleming procedure with Urn allocation. These modifications were introduced to improve the performance and reliability of the procedure when dealing with multiple primary critical values. By using Urn allocation, the procedure is tailored to address specific considerations and optimize the decision-making process during the clinical trial.

Overall, using multiple critical endpoints and modifying the O’Brien and Fleming procedure with Urn allocation are strategies employed to enhance the efficiency of clinical trials. These advancements aim to provide more robust and informative results, ultimately contributing to advancing medical research and patient care.

The decision to utilize Urn’s allocation in this study was motivated by ensuring that more subjects receive the most effective treatment. Health outcomes are maximized, equity is promoted, resources are optimized, and evidence-based practices are advanced by allocating effective treatment in clinical trials. The urn design is a popular adaptive design that has been extensively researched. Such designs balance complete randomization and
perfect balance in treatment assignments, thereby reducing experimental bias. By forcing a small trial to be balanced, the urn design becomes increasingly randomized as the trial size increases. As a result, it is less susceptible to experimental bias than other restricted randomization procedures. In clinical trials, assuming that the study subjects represent a random sample from a uniform population may be challenging. A randomization model provides a preferred basis for statistical inference in such instances.

In any new statistical procedure, the reduction in type I error is crucial as it ensures that false positive results are minimized, thereby enhancing the reliability of the findings. At the same time, maintaining power is essential to ensure that the study has sufficient sensitivity to detect true treatment effects.

To assess the performance of the proposed procedure in controlling type I error, a detailed analysis was conducted by considering alpha = 0.05 and 0.01 and various sample sizes.

In the first case, an alpha value of 0.05 and a sample size of 360 were employed. The investigation began by calculating the type I error, where the results were promising since the type I errors did not exceed the value 0.050808, and most of the values were under 0.05. To further investigate the performance of the proposed procedure, additional analyses were conducted to calculate the type I error rates using other different sample sizes. The type I error values were below the 0.052 threshold.

In addition to the previous analyses, further investigations were conducted to assess the type I error values with an alpha level of 0.01. All type I error values in all cases were below the specified threshold of 0.011, and most of the values were under 0.01. Indicating successful control of type I errors. These findings emphasize the effectiveness of the proposed procedure in controlling type I error. Since the calculated type I error rates consistently fell within acceptable ranges and most of them were less than the desired alpha, affirming the reliability and validity of the modified procedures. By employing Urn’s allocation, the new procedure offered improved control over type I errors, enhancing the integrity of clinical trial outcomes and decision-making processes.

The type I error results were anticipated because, in any study with multiple endpoints, critical values are adjusted to maintain the overall type I error rate. Type I errors occur when a null hypothesis is rejected even though it is true. Adjusting the critical values makes it harder to reject the null hypothesis, reducing the chance of a type I error when $H_0$ is true.

Furthermore, multiple comparisons were made using different success rates and sample sizes to determine if the proposed implementation maintained an acceptable acceptance rate (power) of the alternative hypothesis $H_a$. The study considered success rates of 0.1, 0.15, 0.2, 0.25, and 0.3, with corresponding sample sizes to ensure a power of approximately 0.8 using the Chi-square test. Despite observing a slight decrease in the values, the new implementation successfully maintained the acceptance rate under $H_0$ as being true (power), where the power values were greater than 0.7723 for alpha level 0.05 and 0.78067 for alpha level 0.01—confirming the maintenance of satisfactory power levels.

These results demonstrate that, despite slight decreases, the proposed implementation effectively maintains acceptable power values for alpha levels 0.05 and 0.01 across various test sizes and sample sizes. The observed differences between the values remain within reasonable margins, indicating the reliability and robustness of the modified procedures.

In conclusion, the results of this study suggest that the UMP is a potentially more effective procedure compared to O’Brien and Fleming’s multiple tests and the single sample method. The UMP offers improved control over type I error while maintaining acceptable power levels.

In conclusion, the results of this study highlight the effectiveness of employing multiple critical endpoints in the O’Brien and Fleming procedure, while modifying it using Urn allocation to enhance clinical trials efficiency. The modification introduced in this study successfully controlled type I errors, ensuring reliability and validity of the findings. By utilizing Urn allocation, the procedure addressed specific considerations and optimized the decision-making process during the trial, leading to improved treatment allocation.
and patient outcomes. The proposed implementation maintained an acceptable power level showing its effectiveness in detecting true treatment effects. These findings provide valuable insights into the potential of modified procedures and open avenues for future research.

Future work will compare the UMP with other binary outcome testing methods, particularly the Optimal Weighted Multiple Testing Procedure (OWMP) discussed earlier [39]. This comparison will provide further insights into these different procedures’ relative strengths and weaknesses. Furthermore, the proposed procedure will be extended to more than two treatments. This extension will involve conducting additional iterations based on the modifications applied to the original O’Brien and Fleming procedure. By expanding the scope of the study, a more comprehensive evaluation of the UMP’s performance and applicability can be achieved. By advancing statistical procedures in clinical trials, we can contribute to medical research advancement and ultimately improve patient care.


Funding: This research received funds from the Deanship of Scientific Research at Jordan University of Science and Technology, research grant no. 20220334.

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

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

References
32. Zhao, W.; Weng, Y. Block urn design—A new randomization algorithm for sequential trials with two or more treatments and balanced or unbalanced allocation. *Contemp. Clin. Trials* **2011**, *32*, 953–961. [CrossRef]
35. Neyman, J.; Pearson, E.S. Further notes on the $x^2$ distribution. *Biometrika* **1931**, *22*, 298–305. [CrossRef]
37. Fix, E. *Tables of the Noncentral $x^2$*; University of California: Riverside, CA, USA, 1949; Volume 1, pp. 15–19.
42. Ivanova, A. A play–the-winner type urn model with reduced variability. *Metrika* **2003**, *58*, 1–13. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.