Identification of the Maximum Safe Dose for Binary Endpoints

Abstract

In most clinical trials, binary outcomes—such as success or failure and presence or absence of side effects—are used to evaluate treatment safety and efficacy. Determining the Maximum Safe Dose (MSD) is essential, as it identifies the highest drug dosage that does not cause harmful effects. Exceeding the MSD can pose serious health risks and undermine the overall benefits of the treatment. This article proposed a confidence interval procedure designed to simplify the complex analysis of binary endpoints as discussed in Thall et al. (2008). To solve this problem, we apply a confidence interval process along with a partitioning approach. therefore, application of reliable statistical approaches in establishing and confirming the range of safe dosages is imperative. It involves thorough examination of data in preclinical as well as clinical trials in the hopes of reducing side effects and optimizing the efficacy of treatment. Thus, our paper introduces a confidence interval approach for estimating MSDs for drugs using binary endpoints. This method was performed using a 100(1 - α)% Wilson (1927) score interval with step-up method for binary endpoints without multiplicity adjustment. We illustrate it through the examples which were published by Neuhäuser and Hothorn(1997) in their paper. Additionally, we also observed that this method's power increases with increasing sample size. Finally, our simulation results shows that Wilson score interval proved to have the shortest length and producing good coverage probability. The results further showed that our newly constructed procedure control the familywise error rate. We advocate that our newly constructed procedure with wilson score interval is suitable for demonstrating MSD when the respose have binary outcomes.

keywords: Family-wise error rate, Coverage probability, Binary outcomes, Wilson score interval, confidence-based procedure, Multiple treatments and Confidence-based procedure.

The Authors

John Ayuekanbey Awaab, Mphil., is preparing a PhD dissertation under the supervision of the second author, at Department of Statistics and Actuarial Science, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana.

Michael Jackson Adjabui, Ph.D., is a Senior Lecturer of Mathematics at Department of Mathematics, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana.

Jakperik Dioggban, Ph.D, is a Senior Lecturer of Statistics at Department of Biometry, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana.

1 Introduction

In the vast majority of clinical trials, binary data—that report outcomes which are two types, such as success/failure, side effects/absence of side effects, or survival/death—are used to assess both safety and efficacy of treatments. Dichotomous outcomes necessitate the careful determination of the Maximum Safe Dose (MSD), i.e., the highest degree of dosage of a drug that is not associated with unacceptable adverse effects. It is important to establish the MSD because exceeding this level could result in severe health risks to patients and devalue the general therapeutic benefit of the intervention. Thall and Wathen (2007) and Bauer, and Kohne, (1994) discuss MSD for binary data. They highlighted the importance of identifying the dose at which the risk of adverse effects or disease occurrence is minimized while still achieving the desired therapeutic effect. They emphasize that when the response is binary and the parameter of interest is the risk of the disease, namely the probability of the occurrence of a disease. However, this study can only be found in various biometric studies. Based on this situation, Multiple inference procedures founded on the t-statistic may not be conceivable. This challenge is one of the outstanding questions in the development of multiple comparison strategies for dose finding, according to Tamhane and Dunnett (1999, p. 67). Technically speaking, the challenge is in the strong control of the familywise error rate, which is a crucial factor in limiting the error of mistakenly classifying any of the dangerous dosages as safe.

Iwanami (2001) and Riggs *et al.* (1991) suggests that any drug has the potential to kill when delivered in large enough doses and that all toxicants have a no-observed-effect threshold below which no change in structure or function can be observed. However, drug-related fatalities in Ghana reached 630 in 2020, or 0.36% of all deaths, according to WHO (2018) data given at the time. The age-adjusted mortality rate of 3.31 per 100,000 people places Ghana as the 22nd most populous country in the world. This is because the general public consumes dosages that are hazardous and for this reason, new MSD procedures are needed to enhance the effectiveness of decision and dosage estimation methods, particularly for binary endpoints, in controlling unsafe drug consumption.

Binary endpoints are commonly used in health sectors, particularly cardiovascular and oncology studies, due to

their three main benefits: eliminating the need for multiplicity adjustment, providing a better justification for treatment group distinctions, and increasing power as recorded occurrences increase. These endpoints are crucial as they provide more detailed information about the disease's course (Bretz et al.,2010). Based on this assertion, the binary endpoints are important.

Our paper's goal is to offer stepwise confidence intervals for the identification of MSD for binary endpoints data sets where the proportion of treatment vs control is studied. We employed the partitioning concept, the suggested procedure extends the method proposed by Hsu and Berger (1999) into a stepwise confidence-based procedure without multiplicity adjustment for binary endpoints. The article is organized as follows. In Section 2, the material and methods notations are defined, Our proposed stepwise confidence intervals methods in Section 2.1, the formulation of the proposition 1 in section 3.1, the Family-wise error rate in section 3.2, Results and discussion in section 4, Simulation study of coverage probability, abberration and power comprison and discussion of results in section 4.2. Finally, discussion of results in section 4.5 and conclusion in Section 5.

2 Material and Methods

Suppose that the proportion of k treatments and zero control group called placebo are investigated. Denote $i = 0, 1, 2, \ldots, k$ be a set of inceasesing doses levels used in the study of dose-findings where 0 represents placebo. Consider binary endpoints setting in which a random sample $Y_{i1}, Y_{i2}, Y_{i3}, \ldots, Y_{in_i}$ is the observed response of toxicity from the i^{the} dose level ($i = 0, 1, \ldots, k$). Let P_i be the proportions of toxicity at dose i and P_0 to be proportion of zero control dose such that $P_i = \{P_0, P_1, \ldots, P_k\}$,$i = 0, 1, \ldots, k$ respectively. We assumed that large values of the proportions of treatment represents high toxicity relative to the proportions of the placebo Let $Y_{ij} \sim B(n_i, p_i)$ and $Y_{0j} \sim B(n_0, p_0)$ to be two independent binomial random variables for any $i = 1, \ldots, k$ and $j = 1, \ldots, n_i, n_0$.

In this article, we provide a procedure that will estimate the difference between two unknown proportions $\lambda_{i,0} = P_i - P_0$, $i = 0, 1, \ldots, k$ without loss of generality, let $\lambda_{i,0} = P_i - P_0$, $i = 1, 2, \ldots, k$ be difference of two proportions of interest. Let γ represents the pre-specified threshold constant for toxicity of a drug.The problem of identifying the MSD is formulated as follows;

$$H_{0i} : \bigcup_{i}^{k} \lambda_{i,0} \geq \gamma \text{ versus } H_{ai} : \bigcap_{i}^{k} \lambda_{i,0} < \gamma, i = 1, \ldots, k \tag{1}$$

Assume the random variable Y_{ij} has a distribution determined by the parameter $\theta = (\lambda_{1,0}, \ldots, \lambda_{k,0})$ with $\theta \in \Theta$, where Θ is the parameter space, let $\Theta_i = (\theta : \lambda_{i,0} \geq \gamma)$ and $\Theta_i^c = (\theta : \lambda_{i,0} < \gamma))$ $i = 1, \ldots, k$.

We defined a confidence set, $C(Y)$, for θ is said to be directed toward a subset of the parameter space, $\Theta^* \subset \Theta$ if, for every sample point Y, either $\Theta^* \subset C(Y)$ or $C(Y) \subset \Theta^*$. For the case of one sided significant difference of two proportions inference, say $\Theta_i^c = \lambda_{i,0} < \gamma$, confidence intervals for $\lambda_{i,0}$ of the form $C_i(Y) = (-\infty, W_i(Y))$

are directed towards Θ_i^c for $i = 1, \ldots, k$ and $C_i(Y)$ is the upper limit of the confidence interval. If $D_i(Y)$ is a $100(1-\alpha)\%$ confidence set for $\lambda_{i,0}$, then a $100(1-\alpha)\%$ confidence set, which is directed toward $\lambda_{i,0} < \gamma$, for $\lambda_{i,0}$ is

$$C_i(Y) = \begin{cases} D_i(Y) & \text{if } D_i(Y) \subset \Theta^c \\ D_i(Y) \cup \Theta^c & \text{otherwise} \end{cases} . \tag{2}$$

We create disjoint sets in the parameter space, ensuring that exactly one partition contains the true parameter θ. By achieving this, our procedure effectively controls the Family-Wise Error Rate (FWER) by managing it within each $\Theta_k^* \subseteq \Theta$ for every $k \in K$,where, K is an index. This partitioning strategy provides a robust validation framework, ensuring the accuracy of our inferences while controlling the overall error rate within specified subsets of the parameter space.

2.1 Our Proposed Method

2.2 The proposed method: A Stepwise Procedure

Making inference of MSD with dichotomous responses, we offer a stepwise simultaneous inference approach in this section. The next part provides discussions on the proposition and family-wise error rate control using any standard confidence technique for the difference between two binomial proportions. The first step in solving the problem(1) is to defined the MSD as MSD=max(i;$\lambda_{i,0} < \gamma$). We then employed Wilson (1927) method to obtained the individual $100(1-\alpha)\%$ confidence intervals $C_i(Y)$ for $\lambda_{i,0}$.Taking $C_i(Y)$ and $\lambda_{i,0}$, $i = 1, 2, \ldots, k$ as sample statistics , we test $H_0 = \cup_{i=1}^k \Theta_i$ versus $H_a = \cap_{i=1}^k \Theta_i^c$;

$$W_i(Y) = \frac{1}{(n_i + z_\alpha^2)}((2n_i\lambda_{i,0} - z_\alpha^2) + z_\alpha \sqrt{4n_i\lambda_{i,0}(1-\lambda_{i,0}) + z_\alpha^2})) \tag{3}$$

In the aforementioned expression, $W_i(Y)$ is the upper bound of the confidence interval, n_i is the sample size of dose i and $P(Z \geq z) = 1 - \alpha$ for the standard normal random variable Z, λ_{i0} or the proportion of the risk difference between dose i and zero control dosage group and where $P_i = \frac{y_i}{n_i}$, $P_0 = \frac{y_0}{n_0}$, y_i and y_0 are realizations of the random variables Y_i and Y_0..

2.3 Proposed Procedure

The Proposed Stepwise confidence interval Procedure

Step 1: We identify MSD firstly by computing the upper confidence intervals ($W_i(Y)$) using equation(3.3) for all doses level $i = 1, 2, \ldots k$.

Step 2:We screen the doses in ascending order ($i = 1, \ldots, k$) and designate the dose at B as the MSD such that $W_B(Y) < \gamma$ and $W_{B+1}(Y) \geq \gamma$.

Step 3:However, if $W_1(Y) \geq \gamma$ then no dose is safe and if $W_k(Y) < \gamma$ all dose are safe.

Step 4:We defined the stopping point of the procedure to be step B if $W_B(Y) \leq \gamma$ and $W_{B+1}(Y) \geq \gamma$. So computing confidence intervals for a dose at steps B+1 and B+2 are uneccessary.

Therefore, patients should be strictly limited to doses below "k", which is a preventive measure that reduces the likelihood of exceeding this limit. This positive feedback will reduce the likelihood that patients will suffer unnecessarily from new drugs or drug combinations throughout clinical trials.

3 Our main theoretical results

3.1 Our Proposition

Proposition 1

Let Y represent the set of observed data collected from an experimental study, while Θ defines the parameter space for the parameter vector $\theta = (\lambda_{1,0}, \ldots, \lambda_{k,0})$. Let $C_i(Y)$ be the 100(1-α)% upper confidence limits for $\lambda_{i,0} = P_i - P_0$ for each $i = 1, 2, \ldots, k$ and $\Theta_i^c = \{\lambda_{i,0} < \gamma\}$. Suppose that the procedure stops at step B, where B is the largest integer i such that $C_i(Y) \not\subset \Theta_i^c$,if such an i($1 \leq i \leq k+1$) exists,otherwise, $\Theta_0 = \varnothing$(so $\Theta_0^c = \Theta$) and

$C^*(Y) = \Theta_1^c \cap \Theta_2^c \cap, \ldots, \cap \Theta_{B-1}^c \cap \Theta_B^c \cap C_B(Y)$

For all $\theta \in \Theta$,

$P(\theta \in C^*(Y)) \geq 1 - \alpha$.

Proof of Proposition 1

We denote $\Theta_i^c = \{\lambda_{i,0} < \gamma\}$, and $\Theta_i = \{\lambda_{i,0} \geq \gamma)\}$, $i = 1, \ldots, k$.

We partition the parameter space as follows:

$\Theta_1^* = \Theta_1$

$\Theta_2^* = \Theta_1^c \cap \Theta_2$

$\Theta_3^* = \Theta_1^c \cap \Theta_2^c \cap \Theta_3$

$\vdots$

$\Theta_{k+1}^* = \Theta_1^c \cap \Theta_2^c \cap \Theta_3^c \cap, \ldots, \cap \Theta_k^c \cap \Theta_{k+1}^*$

Since the response are observed over safety across the increasing dose levels.

we now have

$C(Y) = \bigcup_{i=1}^{k+1} C_i(Y) \cap \Theta_i^*$, $\Theta = \bigcup_{i=1}^{k+1} \Theta_i = \bigcup_{i=1}^{k+1} \Theta_i^*$. if $\theta \in \Theta_i^*$, then

The proposition is prove using the the following properties below:

(1).$C_i(Y) \cap \Theta_i^* = \varnothing$ for all $i < B$, because $\Theta_i^* \subset \Theta_i$;

(2).$\Theta_1^c \cap \Theta_2^c \cap, \ldots, \cap \Theta_{B-1}^c \cap C_B(Y) \cap, \ldots, \cap \Theta_{i-1}^c \Theta_i \subset \Theta_1^c \cap \Theta_2^c \cap, \ldots, \cap \Theta_{B-1}^c \cap C_B(Y)$ for all $i > B$.;

(3).$\Theta_B^c \subset C_B(Y)$.

Since the response are observed over safety across the increasing dose levels.

we now have

$$
\begin{aligned}
C^*(Y) &= \bigcup_{i=1}^{k+1} C_i(Y) \cap \Theta_i^* \\
&= \bigcup_{i=1}^{B-1} \bigcup_{i=B}^{B} \bigcup_{i=B+1}^{k+1} C_i(Y) \cap \Theta_i^* \\
&= \bigcup_{i=B}^{B} \bigcup_{i=B+1}^{k+1} C_i(Y) \cap \Theta_i^* \quad \text{(Property 1)} \\
&= \left(C_B(Y) \cap \Theta_B^*\right) \bigcup \bigcup_{i=B+1}^{k+1} C_i(Y) \cap \Theta_i^* \\
&= (\Theta_1^c \cap \Theta_2^c \cap \cdots \cap \Theta_{B-1}^c \cap \Theta_B^c) \bigcup \left(\bigcup_{i=B+1}^{k+1} C_i(Y) \cap \Theta_i^* \right) \\
&\subset \left(\Theta_1^c \cap \Theta_2^c \cap \cdots \cap \Theta_{B-1}^c \cap C_B(Y)\right) \bigcup \left(\Theta_1^c \cap \Theta_2^c \cap \cdots \cap \Theta_{B-1}^c \cap \Theta_B^c\right) \quad \text{(by Property 2)} \\
&= \Theta_1^c \cap \Theta_2^c \cap \cdots \cap \Theta_{B-1}^c \cap \Theta_B^c \cap C_B(Y) \quad \text{(by Property 3)} \\
&= \left(\bigcap_{i=1}^{B} \{\lambda_{i,0} < \gamma\} \right) \cap \{\lambda_{B-1,0} < \gamma\} \cap C_B(Y)
\end{aligned} \tag{4}
$$

We also use this property to have the above results,

$(C_B(Y)\Theta_B) \bigcup (C_B(Y)\Theta_B^c) = (C_B(Y)\Theta_B) \bigcup \Theta_B^c = C_B(Y)$ since $\Theta_B^c \subset C_B(Y)$.

Hence, we have for all $\theta \in \Theta$

$C^*(Y) = \Theta_1^c \cap \Theta_2^c \cap, \ldots, \cap \Theta_{B-1}^c \cap \Theta_B^c \cap C_B(Y))$

$P(\theta \in C^*(Y)) \geq 1 - \alpha$

3.2 Family-wise error rate

Proposition 2 :Stepwise confidence intervals procedure for binary endpoints strongly controls the familywise error rate(FWER) at level α

Proof of Propositon 2

We made the assumption that our procedure ends at step B in order to demonstrate that it controls the FWER. We also proved that our method works for step $B + 1$.

We indicate the composite confidence set up to step B with $C^*(Y)$. Next we obtain: $P(\theta \in C^*(Y)) \geq 1 - \alpha$.

Now that step $B + 1$ has been considered the composite confidence set is still unchanged and the property holds for which $C_{B+1}(Y) \in \Theta^c_{B+1}$.

Thus control of the FWER at level α is ensured since by induction the property holds for all steps up to $k + 1$.Thus the FWER is effectively controlled by the closed Testing procedure which guarantees that for all $\theta \in \Theta$ the probability of θ being in $C^*(Y)$ is at least 1- α.

4 Results and Discussion

4.1 Practical Application of MSD

The study validated theoretical findings by using the lung tumor data set from Neuhäuser and Hothorn's 1997 publication to implement a confidence set-based procedure for determining the MSD. The lung tumor results from 1,2-dichloroethane research show unbalanced sample sizes, typical in carcinogenicity research. The control group was compared with experimental treatments to determine safety for individual dose levels, ensuring control of the family-wise error rate. This approach allows for dichotomous outcomes in toxicological investigations and clinically relevant cut-off values. The study used a binomial distribution random variable to represent a dichotomous endpoint, with a negative control group and various dosage groups. A safety margin of $\gamma = 0.80$ was set to demonstrate the dose's safety. R software was used to derive upper confidence intervals, and the control group was compared to each experimental treatment to control family-wise error rates. Where, UCB denote upper confident bound.

MSD is correctly specified,if and only if $C_i(Y) < \gamma$ and $C_{i+1}(Y) \not< \gamma$, $i = 1, 2, \ldots, k$

Step 1 $C_1(X) = (-\infty, 0.7286) \subset (-\infty, 0.80)$ we reject H_{01} and proclaim that dose 1 is safe and go to step 2.

Step 2 $C_2(X) = (-\infty, 0.9275) \subset (-\infty, 0.80)$ we do not reject H_{02} and proclaim that dose 2 is not safe and stop.

The stepwise confidence interval procedure concludes after step 2, requiring no further testing. Our analysis indicates that dose 1 is deemed safe, while doses 2 is deemed unsafe at a significance level of 0.05. Consequently, we recommend dose 1 as MSD.

Table 1: Upper confidence intervals for Lung tumor data

Dose	Tumor absent	Number at risk	Risk difference	95% UCB
0	35	37	-	-
1	41	48	$P_1 - P_0$	0.7283
2	21	36	$P_2 - P_0$	0.9275

4.2 Simulation Study

Since the binomial distribution exhibited poor coverage probability then we proceeded to investigate the coverage probability of the eleven confidence methods and abberration in table 2 and 3 in order to select an appropriate confidence interval that guarantee good coverage probability and was free from abberration problem. We also estimated the length of the various confidence interval methods in order to determine which of the methods that are the most efficient in estimating confidence intervals for binary outcomes. We measured the efficiency of the confidence methods by taking confidence method that has the shortest length. However, we conducted 10,000 simulations to assess the performance of eleven confidence interval methods in terms of coverage probability, length and abberration. The reason was that any confidence interval methods that ensures good coverage probability and free from abberration are considered in dose findings. We considered the various risk proportions and sample sizes as (n_i = 20, 18, 10, 48, and 36, for $i = 1, 2, \ldots, 5$), incorporating differences between two binomial proportions ($P_1 - P_0 = 0.25$, $P_2 - P_0 = 0.51$, $P_3 - P_0 = 0.50$, $P_4 - P_0 = 0.1$, and $P_5 - P_0 = 0.37$). The nominal confidence level was set at 90%. Subsequently, we computed the coverage probability for each of the eleven conditional confidence interval methods. A detailed presentation of the results is provided in Table 2 below: As per the findings presented in Table 2, the prop test, Bayes, logit, and exact confidence intervals exhibited inadequate coverage and disproportionately long lengths. In contrast, the remaining confidence techniques demonstrated commendable coverage probability and relatively shorter lengths. Notably, the Wilson confidence interval outperformed its counterparts, boasting the smallest length among them. General observations from the simulation exercise consistently indicated that the Wilson score method exhibited a mean coverage probability that was not only shorter but also closer to the nominal level. This implies that methods producing confidence intervals with robust coverage probabilities and relatively shorter lengths are more effective.

4.3 Comprison of FWER

To investigate the familywise error rate performance of the three techniques, we ran 10,000 simulations. For the difference between two independent binomial proportions, we used a step-up technique with different confidence interval methods like Wald confidence, Wald CC confidence, and Wilson's score interval. We ran simulations un-

Table 2: Comprison of wilson score interval with other confidence methods under coverage probability

Methods	CP(Length) n_1=20 P1-P0=0.25	CP(Length) n_2=18 $P_2 - P_0$=0.51	CP(Length) n_3=10 P3-P0=0.50	CP(Length) n_4=48 $P_4 - P_0$=0.10	CP(Length) n_5=36 $P_5 - P_0$=0.37
Agresti-C	0.9353(0.3028)	0.9058(0.3537)	0.8901(0.4463)	0.9153(0.1478)	0.8869(0.2521)
Asymp	0.8668(0.3068)	0.9058(0.3765)	0.8901(0.4911)	0.8500(0.1376)	0.8869(0.2593)
Bayes	0.8668(0.2912)	0.9058(0.3576)	0.8901(0.4480)	0.8232(0.1332)	0.8869(0.2524)
cloglog	0.8668(0.2958)	0.9424(0.3662)	0.9314(0.4695)	0.9153(0.1373)	0.9156(0.2539)
Exact	0.9353(0.3415)	0.9424(0.4072)	0.9774(0.5317)	0.9153(0.1373)	0.8869(0.2544)
logit	0.9353(0.3057)	0.9119(0.4075)	0.8901(0.4603)	0.9512(0.1460)	0.8869(0.2547)
Irt	0.8668(0.2991)	0.9058(0.3607)	0.8901(0.4618)	0.9153(0.1376)	0.8869(0.2549)
Probit	0.9353(0.3022)	0.9058(0.3631)	0.8901(0.4632)	0.9153(0.1413)	0.8869(0.2547)
Profile	0.8668(0.2992)	0.058(0.3631)	0.8901(0.4618)	0.9153(0.1376)	0.8869(0.2549)
Prop.test	0.9887(0.3883)	0.9692(0.4462)	0.9774(0.5394)	0.9780(0.1888)	0.9754(0.3196)
Wilson Sc	0.9353(0.2961)	0.9058(0.3525)	0.9018(0.2988)	0.9153(0.1414)	0.9088(0.2511)

der one-sided ($\alpha = 0.025$) and two-sided ($\alpha = 0.05$) conditions to compare the familywise error rate performance of the three techniques. The steps' specifics are provided below:

Procedure A: Step-up method(without adjusting the alpha level) with Wilson's score interval

Procedure B: Step-up method(without adjusting the alpha level) with Wald's confidence interval

Procedure C: Step-up method(without adjusting the alpha level) with Wald's CC confidence interval

The FWER for three distinct techniques (A, B, and C) at two significance levels ($\alpha = 0.025$ and $\alpha = 0.050$) are shown in Table 4 for a range of sample sizes (n). Procedure A's FWER falls between 0.0103 and 0.0237. At $n = 15$ (0.0103), the lowest FWER is recorded, and at $n = 45$ (0.0237), the greatest. This process keeps the FWER low and comparatively constant for varying sample sizes.Procedure B's FWER ranges from 0.0084 to 0.0269. At $n = 25$ (0.0084), the lowest FWER is recorded, and at $n = 75$ (0.0269), the highest. Procedure B too exhibits erratic performance, albeit not significantly.Procedure C's FWER falls between 0.0084 and 0.0265. Similar to Procedure B, the lowest FWER is seen at $n = 25$ (0.0084), and the greatest is at $n = 65$ (0.0265). In terms of continuing to have a high error rate, Procedure C shows a similar tendency to Procedure B. It is clear that some of the FWER values for methods B and C are either much lower or much higher than the nominal value, indicating inadequate control over the familywise error rate. Table 3 clearly shows that process A performs better than procedures B and C at $\alpha = 0.025$ when comparing the FWER performance of the three procedures, as some of the FWER values for procedures B and C are comparatively higher than the nominal value. Nevertheless, Procedure A appears to perform better than B and C at 0.050 in terms of keeping FWER around or below the nominal threshold. According to the results, Procedure A is recommended since it ensures that the FWER for both one-sided and two-sided situations is under control.Procedure A is the recommended method since it exhibits better control over the FWER at both significance levels. Even if they work well, procedures B and C show more variability and sporadic peaks in FWER, particularly for $n = 15$ for $\alpha = 0.05$. Because Procedure A consistently keeps the FWER near to or below the nominal level for both one-sided and two-sided testing, it is therefore advised.

Table 3: Comparison of Familywise error rate among the three procedures

	$\alpha = 0.025$			$\alpha = 0.050$		
n	Procedure A	Procedure B	Procedure C	Procedure A	Procedure B	Procedure C
10	0.0194	0.0206	0.0185	0.0195	0.0197	0.0185
15	0.0103	0.0125	0.0092	0.0105	0.0743	0.0670
25	0.0217	0.0084	0.0084	0.0401	0.0422	0.0363
35	0.0174	0.0194	0.0179	0.0190	0.0572	0.0550
45	0.0237	0.0241	0.0226	0.0236	0.0663	0.0598
55	0.0201	0.0261	0.0245	0.0235	0.0582	0.0588
65	0.0107	0.0254	0.0265	0.0265	0.0623	0.0590
75	0.0228	0.0269	0.0234	0.0249	0.0501	0.0504
85	0.0231	0.0220	0.0221	0.0419	0.0453	0.0425
95	0.0195	0.0155	0.0227	0.0376	0.0376	0.0419
100	0.0186	0.0204	0.0173	0.0346	0.0401	0.0342

4.4 Comparison of Powers

We conducted 10000 simulation to study the power performance of the three procedures. We compared a step-up technique for the difference between two independent binomial proportions with a method based on Wald confidence, Wald CC confidence, and Wilson's score. The family-wise error rate is managed by the step-up approach without multiplicity correction, which does not take the population distribution into account. We just consider the support method in our evaluation; no other assumptions are made. We take into consideration the next three comparisons in this context.

Procedure A: Step-up method(without adjusting the alpha level) with Wilson's score interval

Procedure B: Step-up method(without adjusting the alpha level) with Wald's confidence interval

Procedure C: Step-up method(without adjusting the alpha level) with Wald's CC confidence interval

As can be seen in Table 4, the power of procedure A increases as the sample size increases, reaching 100 (or 100% power) for sample sizes of 40 and above. This shows that procedure A is effective in statistical analysis even with a small sample size (e.g., the power of 0.9965 at n = 25). The performance of Procedure A shows that it is a reliable method to obtain the best results for simulation. The threshold value for a sample size of 5 is 40 and for a sample size of 100 it is as low as 0.546, with the power increasing to 0.701. This is 0.533 when n = 30 and 0.456 when n = 35, showing some discrepancies. In general, procedure B was found to be valid but inferior to Method A in terms of consistency and reliability. It reached 0.799 at n = 40 and then changed slightly. The power changes very little at the average rate (e.g. the power decreases to 0.488 for n = 35, to 0.601 for n = 70). Although procedure C does not achieve the same power as A, it is generally similar to B and shows the performance of different models. This shows that it is quite sensitive and reliable in detecting the effects in simulations. These changes indicate that the method may not always be reliable, especially in small samples. It improves with increasing volume and can be a valid alternative when procedure A is not appropriate. This is desirable because larger samples can increase the information and power of the test. Procedure A demonstrates that even a small sample size (n = 25 to n = 40)

is sufficient to obtain high power, whereas procedure B and C may require larger samples to obtain comparable power. It is the most efficient and is preferred if high sensitivity is desired. procedure C provides greater energy efficiency and can be considered where procedure A cannot be applied, especially if it is important to maintain a certain energy level between different models.

Table 4: Comprison of Powers

Sample size	Power of A	Power of B	Power of C
5	0.7357	0.546	0.675
10	0.7441	0.658	0.688
15	0.9495	0.648	0.679
20	0.9914	0.605	0.689
25	0.9965	0.659	0.701
30	0.9996	0.533	0.635
35	0.9999	0.456	0.488
40	1.0000	0.701	0.799
50	1.0000	0.688	0.691
70	1.0000	0.599	0.601
100	1.0000	0.701	0.771

4.5 Discussion of Results

In situations where safety margins were set at 80%, we employed meticulous stepwise confidence procedures to calculate the MSD of lung tumor data, resulting in an MSD of dose 1 for binary data. This determination holds paramount significance for both pharmaceuticals and consumers. Given that a drug's adverse effects intensify with dosage, and patient concerns predominantly revolve around safety rather than efficacy, our focus was directed towards pioneering novel stepwise confidence-based procedures to pinpoint the MSD in binary endpoints. The procedures elucidated in this study exhibit robust control over the familywise error rate, and their validation through the partition principle bolsters their credibility. The significance of coverage probability cannot be overstated in clinical dose-finding studies, as it mitigates the risk of erroneously labeling an unsafe dose as safe. Neglecting this aspect could have adverse consequences for drug users. Consequently, we meticulously assessed the performance of eleven binary confidence interval methods in terms of coverage probability. Our findings reveal that while the asymptotic confidence interval is conventional, it consistently exhibited poor coverage probability throughout the study. In contrast, the prop test, Bayes, Logit, and exact methods consistently demonstrated subpar coverage probabilities. On the other hand, the Wilson score, Agresti, Profile, and LRT methods consistently ensured a high coverage probability. However, we investigated the efficiency level of the various confidence methods by computing the length for each method and the Wilson score method recorded the shortest length throughout the simulation study. Thus, it was evident that the Wilson score interval was the most efficient method for estimating

confidence intervals under binary endpoints. For the case of the power comparison among the three procedures, it was clear from the findings that procedure A recorded the highest performance among its counterparts B and C. Thus, the procedure attained 100% performance at a relatively smaller sample of 40. But when Procedure A isn't applicable, Procedure C in particular can be thought of as a good substitute. Also, when we further compare the FWER performance among the three procedures, it was evident from the findings that procedure A strongly controls the FWER whereas B and C do not. Although Procedures B and C also perform well, they exhibit greater variability and sporadic peaks in FWER, especially for $n = 15$ at $\alpha = 0.05$. Procedure A is the preferred method for controlling the FWER in binary data, and Wilson's score is recognized as the most efficient method for estimating confidence intervals in binary data, based on newly constructed stepwise confidence-based procedures. Wilson's score stands out due to its strong coverage probability and shorter interval lengths. In a dose-response study, the criteria for selecting an appropriate confidence method said to be the method must control the FWER, and guarantee good coverage probability.

5 Conclusion

Based on the findings and discussion of the results, we concluded that our newly constructed confidence-based procedure (A) strongly controlled the FWER and as well as obtained the highest power performance among the three procedures compared. Therefore, we recommend that our newly constructed confidence-based procedure coupled with Wilson's score interval, is the most suitable for estimating confidence intervals for binary endpoint trials.

Data Availability

We make used of the data that was published by Neuhäuser and Hothorn.(1997),Table 1, page 464. Also, see the summary data in table 1 of our article, page 7.

Competing Interests

Authors have declared that no competing interests exist.

References

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation applications. *J. Statist. Plann. Infer*, 82,55-68.

A test of homogeneity for ordered alternatives. Biometrika,46,34-38.

Bauer, P. and Budde, M.(1994).Multiple testing for detecting efficient dose steps. *Biom. J.*, 36, 3-15.

Bauer, P. and Kieser, M. (1996).A unifying approach for confidence intervals and testing of equivalence and difference.*Biometrika*, 83, 934-937.

Bauer, P.(1991).Multiple testing in clinical trials. Statistics in Medicine, 10, 871-890.

Bauer, P., Rohmel, J., Maurer, W. and Hothorn, L. (1998).Testing strategies in multi-dose experiments including active control. *Statistics in Medicine*,17, 2133-2146.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Biometrical Journal*. 50(4), 480-504.

Body, J., Quebe-Fehling, E., and Seaman, J.(2001).Zoledronic acid is superior to pamidronate in the treatment of hypercalcemia of malignancy: A pooled analysis of two randomized, controlled clinical trials. *Journal of Clinical Incology*,19,(2),558-567.

Bornkamp,B., Bretz, F., Dette, H. and Pinheiro, J.(2011). Response-adaptive dose-finding under model uncertainty. *Ann Appl Stat.*,5(2B),1611-1631.

Bretz, F., Dette, H. and Pinheiro, J. (2010). Practical considerations for optimal designs in clinical dose finding studies.Stat Med,29, 731-742.

Bretz, F., Pinheiro, J. and Branson M.(2005).Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61,738-748.

Brown, L.D., Cai, T. and Dagupta, A. (2001). Interval estimation for a binomial proportion,*Statistical Science*.**16**,101-133.

Chen,J.(2008).Inference on minimum effective dose with binary data. Technical Report,Department of Mathematics and Statistics, Bowling Green State University.

Chen, J.(2006).Minimum effective dose with binary data. Technical Report, Department of mathematics.

Gupta,S.S. and Berger,J.O.(1995).Independent binomial proportions. J. Statist. *Plann. Infer.*,1,97-115.

Hsu,J. C. and Berger,R.(1999) Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *J. Amer. Statist*.94,468-482.

Hochberg, Y., and Tamhane, A. C.(1987) Multiple Comparison Proce- dures, New York: Wiley.

Iwanami, A.(2001). Psychotropic-induced water intoxication and its countermeasures. *Journal of the Japanese Medical Association*, 44 (9), 417-22.

Nadew, S.S., Beyene, K.G.M. and Beza, S.W.(2020). Adverse drug reaction reporting practice and associated factors among medical doctors in government hospitals in Addis Ababa, Ethiopia.

Newcombe, R.G. (1998a). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17,857-872.

Newcombe, R.G. (1998b). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17, 873-890.

Riggs, A.T.,Dysken, M., Kim, S.W. and Opsahl, J.M.1991. A review of disorders of water homeostatis in psychiatric patients. Psychosomatics, 32(2), 133-46.

Scheffe, H.(1953).A method for judging all contrasts in the analysis of variance. *Biometrika*, 40,87-104.

. Clinical and economic burden of adverse drug reactions.*J Pharmacol Pharmacother*, **4**(1),73-77.

Tallarida,R. J.(2000). Drug Synergism and Dose-Effect Data Analysis. New York: Chapman and Hall/CRC.

Tamhane,A. C. and Dunnett,C.W.(1999) Stepwise multiple test proceudres with biometric applications. *J. Statist. Plann. Infer*.82,55-68.

WHO(2018). Family planning/Contraception. Accessed 27 Dec 2018.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22,209-212.