GRANT journal
ISSN 1805-062X, 1805-0638 (online)
Vol. 13, issue 02
EUROPEAN GRANT PROJECTS | RESULTS | RESEARCH & DEVELOPMENT | SCIENCE

Comparing binomial proportions in clinical trials: exact confidence intervals for small sample sizes

Ivana Pobočíková[1]
Daniela Sršníková[2]
Mária Michalková[3]

[1] Department of Applied Mathematics, Faculty of Mechanical Engineering, University of Žilina, Univerzitná 8215/1, 010 26 Žilina, Slovakia; ivana.pobocikova@fstroj.uniza.sk
[2] Department of Applied Mathematics, Faculty of Mechanical Engineering, University of Žilina, Univerzitná 8215/1, 010 26 Žilina, Slovakia; daniela.srsnikova@fstroj.uniza.sk
[3] Department of Applied Mathematics, Faculty of Mechanical Engineering, University of Žilina, Univerzitná 8215/1, 010 26 Žilina, Slovakia; maria.michalkova@fstroj.uniza.sk

Grant: KEGA projects No. 029ŽU-4/2022 and No. 025ŽU-4/2024
Name of the Grant: KEGA No. 029ŽU-4/2022 Implementation of the principles of blended learning into the teaching of the subject Numerical Methods and Statistics, KEGA No. 025ŽU-4/2024 Implementation of new didactic tools to increase the quality of mathematics teaching in the engineering degree at technical universities
Subject: AM – Pedagogy and education

Abstract The confidence intervals for the difference of two independent binomial proportions are often used in clinical trials to compare a new treatment with a standard treatment. A traditional approach based on standard normal approximation does not work well for a small sample size. This article describes the exact Chang-Zhang and Agresti-Min confidence intervals, which are better alternatives for small sample sizes. Both methods are strictly conservative, ensuring that the minimum coverage probability is always met. We illustrate the use of these intervals with a real example from clinical studies.

Keywords binomial distribution, difference of two proportions, confidence interval, Chan-Zhang interval, Agresti-Min interval, Wald interval, clinical trial

1. INTRODUCTION

In medical studies, assessing the difference between two independent binomial proportions from a comparative study or experiment is often a key research focus. Confidence intervals are an effective way to evaluate this difference. In clinical trials, confidence intervals for the difference of two independent binomial proportions are often used to compare a new treatment with a standard treatment or a placebo or to compare the effects of two drugs. This situation can be illustrated with a 2×2 contingency table.

Table1. Comparison of New Treatment vs. Standard Treatment

	New treatment	**Standard treatment**
Number of successes	X	Y
Number of failures	$n_1 - X$	$n_2 - Y$
Total	X	n_2

Let $X \sim Bi(n_1, \pi_1)$ and $Y \sim Bi(n_2, \pi_2)$ be two independent binomial random variables. Random variable X is the number of successes in the group with a new treatment, π_1 denotes the probability of success, and n_1 is the sample size. Let random variable Y be the number of successes in the group with the standard treatment, π_2 denoting the probability of success, and n_2 is the sample size. The difference between proportions, or the success probabilities, serves as an important effect measure when comparing new and standard treatments. This difference between binomial proportions is denoted as $\delta = \pi_1 - \pi_2$. Apparently $-1 < \delta < 1$. Let $\pi = \pi_1$ and substitute $\pi_2 = \delta - \pi_1$. Then the joint probability mass function can be expressed as

$$P(X = x, Y = y) =$$

$$= \binom{n_1}{x}\binom{n_2}{y}\pi^x(1-\pi)^{n_1-x}(\pi-\delta)^y(1-\pi+\delta)^{n_2-y}$$

for $x = 0, 1, \dots, n_1,\ y = 0, 1, \dots, n_2;\ n_1, n_2 \in \mathcal{N}$ and $\pi_1, \pi_2 \in (0, 1)$. For any given δ the domain of π is

$$D(\delta) = \{\pi: max\{0, \delta\} \le \pi \le min\{1, 1 + \delta\}\}.$$

To evaluate the treatment difference, we aim to find the $100 \times (1-\alpha)\%$ two-sided confidence interval for the difference of two independent binomial proportions. This interval is denoted as $\langle \delta_L, \delta_U \rangle$. The maximum likelihood estimators (MLEs) for the parameters π_1 and π_2 from samples are given by

$$p_1 = \frac{X}{n_1} \text{ and } p_2 = \frac{Y}{n_2},$$

respectively, where X is the number of successes in a random sample of size n_1 and Y is the number of successes in the random sample of size n_2.

The literature offers several methods for constructing confidence intervals for the difference of two independent binomial proportions. This topic has garnered significant attention due to its numerous practical applications. One traditional approach relies on the standard normal approximation. Among these methods, the asymptotic Wald interval is widely employed, and it is defined as

$$\delta_L = (p_1 - p_2) - k_{1-\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}},$$

$$\delta_U = (p_1 - p_2) + k_{1-\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}},$$

where k_α is the α-quantile of standard normal distribution.

In contrast, the exact confidence intervals are derived from exact binomial distribution. These intervals are constructed by reversing a hypothesis test under an appropriate alternative hypothesis. Two commonly used methods are the Chan-Zhang interval (Chan and Zhang, 1999) and the Agresti-Min interval (Agresti and Min, 2001), both will be detailed in the following section.

It is challenging to definitively recommend one method as superior. A critical aspect in evaluating the performance of a confidence interval lies in considering the coverage probability, conservatism, and interval length. These criteria are discussed in greater depth, for instance, in Newcombe (1998).

The coverage probability of the confidence interval $\langle \delta_L, \delta_U \rangle$ is for fixed $n_1, n_2 \in \mathcal{N}$ and $\pi_1, \pi_2 \in (0,1)$ defined by

$$C_{n_1,n_2}(\pi_1,\pi_2) =$$

$$= \sum_{x=0}^{n_1}\sum_{y=0}^{n_2}\binom{n_1}{x}\binom{n_2}{y}{\pi_1}^x(1-\pi_1)^{n_1-x}{\pi_2}^y(1-\pi_2)^{n_2-y}I(x,y,\pi_1,\pi_2),$$

where indicator function $I(x,y,\pi_1,\pi_2) = 1$ if $\delta \in \langle \delta_L, \delta_U \rangle$ and $I(x,y,\pi_1,\pi_2) = 0$ otherwise.

The confidence interval is strictly conservative, if for all $n_1, n_2 \in \mathcal{N}$ and $\pi_1, \pi_2 \in (0,1)$

$$C_{n_1,n_2}(\pi_1,\pi_2) \geq 1-\alpha.$$

The expected length of the confidence interval is defined by

$$EL_{n_1,n_2}(\pi_1,\pi_2) =$$

$$= \sum_{x=0}^{n_1}\sum_{y=0}^{n_2}[\delta_U(x,y) - \delta_L(x,y)]\binom{n_1}{x}\binom{n_2}{y}{\pi_1}^x(1-\pi_1)^{n_1-x}{\pi_2}^y(1 - \pi_2)^{n_2-y}I(x,y,\pi_1,\pi_2),$$

where $\delta_L(x,y), \delta_U(x,y)$ are bounds of a particular confidence interval.

In other words, the coverage probability is the probability that the confidence interval contains the true value. The better confidence interval is such an interval, in which coverage probability is close to the nominal level $(1-\alpha)$. Additionally, shorter intervals are generally preferred because they provide more precise estimates.

In the early phases of clinical trials, the sample sizes are usually small or moderate and strict conservatism is required, so confidence intervals based on large sample approximations do not achieve the nominal level and may not be reliable. While the Wald interval is straightforward to compute, it is widely recognized to perform inadequately for small sample sizes and when the proportions π_1 or π_2 near to boundaries 0 or 1. Various comparisons in the literature, including studies by Newcombe (1998), Agresti and Caffo (2000), and Brown and Li (2005), consistently report poor performance of the Wald interval in terms of coverage probability. Superior alternatives to the asymptotic Wald interval, noted for their improved performance and simplicity of calculation, have been proposed by researchers such as Newcombe (1998), Agresti and Caffo (2000), and Miettinen and Nurminen (1985).

It is known that the exact intervals are strictly conservative, they guarantee the coverage probability above or equal to the nominal level $(1-\alpha)$ and are more reliable when the sample sizes are small or when the proportions π_1 or π_2 are near to the boundaries 0 or 1.

In this paper, we illustrate the application of the confidence intervals for the difference of two independent binomial proportions in clinical trials. We use the real data from the clinical trials. We consider exact Chan-Zhang and the Agresti-Min intervals. Both intervals are strictly conservative and are recommended to be used in clinical trials when strict conservatism is required due to safety and efficacy. The Agresti-Min interval has a coverage probability closer to the nominal level and is less conservative compared to the Chang- Zhang interval. In general, the Agresti-Min interval is shorter than the Chan-Zhang interval (Pobočíková, 2011).

2. EXACT CONFIDENCE INTERVALS

2.1 Chan-Zhang interval

Chan and Zhang (1999) proposed the exact confidence interval by inverting two one-sided score tests

$$H_0\colon \delta = \delta_0 \text{ versus } H_0\colon \delta < \delta_0 \text{ and } H_0\colon \delta = \delta_0 \text{ versus } H_0\colon \delta > \delta_0.$$

and used for testing the score test statistic

$$Z(X,Y,\delta_0) = \frac{p_1 - p_2 - \delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}},$$

where p_1, p_2 are maximum likelihood estimators of π_1, π_2 and $\hat{p}_1, \hat{p}_2$ are maximum likelihood estimators of π_1, π_2 under the restriction that $\hat{p}_1 - \hat{p}_2 = \delta_0$. Miettinen and Nurminen (1985) showed that $\hat{p}_1, \hat{p}_2$ can be obtained uniquely by closed form.

For given $X = x, Y = y$ are the exact one-sided p-values for δ_0 defined by

$$\beta_{CZL}(x,y|Z,\delta_0) =$$

$$= \max_{\pi \in D(\delta_0)}\left\{\sum_{i=0}^{n_1}\sum_{j=0}^{n_2}P(X=i,Y=j|\delta_0,\pi)\, I_1\big(Z(i,j,\delta_0) \geq Z(x,y,\delta_0)\big)\right\},$$

$$\beta_{CZU}(x,y|Z,\delta_0) =$$

$$= \max_{\pi \in D(\delta_0)}\left\{\sum_{i=0}^{n_1}\sum_{j=0}^{n_2}P(X=i,Y=j|\delta_0,\pi)\, I_2\big(Z(i,j,\delta_0) \leq Z(x,y,\delta_0)\big)\right\},$$

where

$$I_1(A \geq B) = \begin{cases}1 & \text{if } A \geq B\\ 0 & \text{otherwise}\end{cases},\quad I_2(A \leq B) = \begin{cases}1 & \text{if } A \leq B\\ 0 & \text{otherwise}\end{cases}$$

are indicator functions.

The $100 \times (1-\alpha)\%$ Chan-Zhang interval is defined by

$$\delta_L = \inf_{\delta}\left\{\delta;\ \beta_{CZL}(x,y|Z,\delta_0) > \frac{\alpha}{2}\right\},$$

$$\delta_U = \sup_{\delta}\left\{\delta;\ \beta_{CZU}(x,y|Z,\delta_0) > \frac{\alpha}{2}\right\}.$$

2.2 Agresti-Min interval

Agresti and Min (2001) proposed the exact confidence interval by inverting one two-sided score test

$$H_0: \delta = \delta_0 \text{ versus } H_0: \delta \neq \delta_0.$$

For given $X = x, Y = y$ is the exact two-sided p-value for δ_0 defined by

$$\beta_{AM}(x,y|Z,\delta_0) =$$

$$\max_{\pi \in D(\delta_0)} \left\{ \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} P(X=i, Y=j|\delta_0,\pi) I(|Z(i,j,\delta_0)| \geq |Z(x,y,\delta_0)|) \right\},$$

where

$$I(|A| \geq |B|) = \begin{cases} 1 & \text{if } |A| \geq |B| \\ 0 & \text{otherwise} \end{cases}$$

is an indicator function.

The $100 \times (1-\alpha)\%$ Agresti-Min interval is defined by

$$\delta_L = \inf_{\delta}\{\delta;\ \beta_{AM}(x,y|Z,\delta_0) > \alpha\},$$

$$\delta_U = \sup_{\delta}\{\delta;\ \beta_{AM}(x,y|Z,\delta_0) > \alpha\}.$$

3. EXAMPLES

In this section, we illustrate the confidence intervals discussed using real clinical study data reported by Gomez-Vasquez et al. (2007) and Mo and Qiu (2017).

3.1 Example 1

Gomez-Vasquez et al. (2007) investigated the efficacy of dexmedetomidine in pain relief after knee arthroscopic surgery.

Dexmedetomidine, an α_2 agonist, is a sedative known for its analgesic properties. Its efficacy and side effects were examined in the immediate postoperative period following knee arthroscopic surgery, a procedure often associated with significant postoperative pain necessitating analgesic intervention. A double-blind, double-placebo-controlled clinical trial was conducted involving 30 patients diagnosed with chronic degenerative knee arthritis or anterior cruciate ligament injury, randomly assigned to receive either intravenous dexmedetomidine or intravenous propacetamol. Pain scores, supplemental analgesic requirements, and side effects were closely monitored. The study aimed to evaluate both substances' efficacy in pain management following knee arthroscopy and their respective side effect profiles. Data regarding the number of patients requiring additional analgesics for pain relief can be visualized using a 2×2 contingency table.

Table 2. Requests for additional analgesics

Request	Dexmedetomidine group	Propacetamol group
Yes	7	4
No	8	11
Sum	15	15

Let π_1 represent the probability of requests for additional analgesics in the dexmedetomidine group, and π_2 denote the corresponding probability in the propacetamol group.

In our study, we have $x = 7, n_1 = 15, y = 4$ a $n_2 = 15$. The maximum likelihood estimators for parameters π_1 and π_2 from the samples are

$$p_1 = \frac{7}{15} = 0.4667 \text{ and } p_2 = \frac{4}{15} = 0.2667.$$

The observed difference is $\delta = p_1 - p_2 = 0.2$. We calculate the 95% confidence intervals for the difference $\delta = \pi_1 - \pi_2$ (difference dexmedetomidine - propacetamol). The results are provided in Table 3.

Due to the limited sample sizes, exact confidence intervals are employed to improve reliability and reduce coverage uncertainties. In clinical trials, ensuring patient safety is of paramount importance, requiring rigorous measures to minimize potential risks.

Table 3. 95% confidence intervals for the difference $\delta = \pi_1 - \pi_2$

Method	$\langle \delta_L, \delta_U \rangle$
Chan-Zhang interval	$\langle -0.1598; 0.5262 \rangle$
Agresti-Min interval	$\langle -0.1506; 0.5181 \rangle$

The methods employed yield consistent sets of confidence intervals, indicating no statistically significant difference between the parameters π_1 and π_2 (intervals include 0). This suggests that there is no significant difference in the analgesic efficacy between the two substances. The intervals show similar widths, with the Chan-Zhang interval being wider than the Agresti-Min interval (Fig. 1).

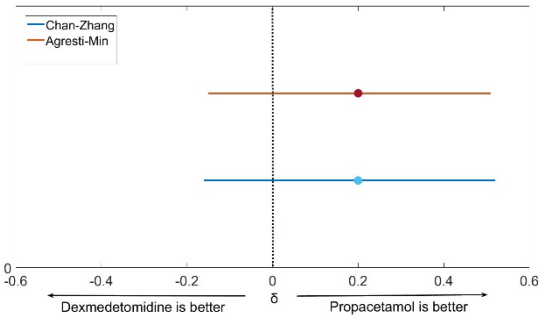

Figure 1. 95% confidence intervals for difference $\delta = \pi_1 - \pi_2$

Throughout the study, adverse effects associated with both substances were carefully monitored in both groups of patients. We will show the interpretation of confidence intervals on selected adverse events. The frequency of specific adverse events is detailed in Table 4, alongside their corresponding 95% two-sided confidence intervals provided in Table 5. Statistical significance is indicated by the symbol *.

Table 4. Adverse effects, number of cases

Adverse event	Dexmedetomidine group	Propacetamol group	Observed difference $p_1 - p_2$
Bradycardia	6	1	0.3333
Hypotension	1	2	-0.0667
Hypertension	5	0	0.3333
Local pain	0	11	-0.7333
Shivering	0	4	-0.2667
Nausea	0	2	-0-1333
Vomiting	0	1	-0,0667
Headache	1	0	0.0667

Table 5. 95% confidence intervals for the difference $\delta = \pi_1 - \pi_2$

Adverse event	Chan-Zhang	Agresti-Min
Bradycardia *	$\langle 0.0117;\ 0.6197\rangle$	$\langle 0.0234;\ 0.6015\ \rangle$
Hypotension	$\langle -0.3439;\ 0.2031\rangle$	$\langle -0.3354;\ 0.2034\rangle$
Hypertension *	$\langle 0.0784;\ 0.6162\rangle$	$\langle 0.0874;\ 0.5939\rangle$
Local pain *	$\langle -0.9221;\ -0.4395\rangle$	$\langle -0.9033;\ -0.4500\rangle$
Shivering *	$\langle -0.5510;\ -0.0093\rangle$	$\langle -0.5290;\ -0.0235\rangle$
Nausea	$\langle -0.4046;\ 0.1014\rangle$	$\langle -0.3923;\ 0.0996\rangle$
Vomiting	$\langle -0.3195;\ 0.1599\rangle$	$\langle -0.3190;\ 0.1513\rangle$
Headache	$\langle -0.1599;\ 0.3195\rangle$	$\langle -0.1513;\ 0.3190\rangle$

The methods consistently produce comparable sets of confidence intervals and agree on their findings regarding adverse effects. In the dexmedetomidine group, notable adverse effects identified were decreased heart rate and high blood pressure. These confidence intervals are statistically significant, as they do not include 0 and have positive endpoints, implying a greater incidence of decreased heart rate or high blood pressure among patients receiving dexmedetomidine (Fig. 2 and Fig. 3).

Conversely, significant adverse effects observed in the propacetamol group included local pain and shivering. The confidence intervals for these effects are also statistically significant, with endpoints that do not include 0 and are negative. This indicates a higher proportion of patients experiencing local pain or shivering with propacetamol (Fig. 2 and Fig. 3).

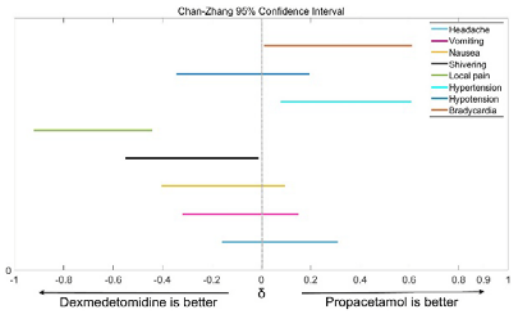

Figure 2. Chan-Zhang 95% confidence intervals for adverse events

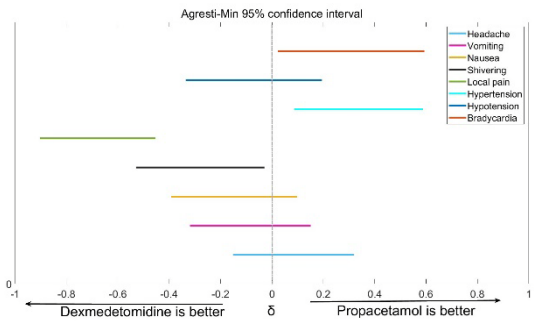

Figure 3. Agresti-Min 95% confidence intervals for adverse events

3.2 Example 2

In the following example, we will show that the Wald interval is not suitable for use when the proportions π_1 or π_2 are near the boundary of 0.

Mo and Qiu (2017) studied the analgesic effect and the impact on adverse reactions of using dexmedetomidine after caesarean section.

Eighty women who underwent caesarean section with combined spinal and epidural anaesthesia were selected for the study. The patients were randomly divided into experimental and control groups, with each group consisting of 40 patients. The patients in the experimental group received ropivacaine hydrochloride and dexmedetomidine, while those in the control group received ropivacaine hydrochloride and morphine. The study focused on the effects of dexmedetomidine in reducing adverse reactions after caesarean section.

Table 6 shows selected adverse events from the study and corresponding 95 % two-sided confidence intervals are listed in Table 7. Statistical significance is indicated by the symbol *.

Table 6. Adverse reactions, number of cases

Adverse reaction	Dexmedetomidine group	Control group	Observed difference $p_1 - p_2$
Nausea	3	15	-0.3
Vomiting	0	6	-0.15
Shakes	2	6	-0.025
Pruritus	2	11	-0.225
Hypotension	0	0	0

Table 7. 95% confidence intervals for the difference $\delta = \pi_1 - \pi_2$

Adverse reaction	Chan-Zhang	Agresti-Min
Nausea *	$\langle -0.4778;\ -0.1037\rangle$	$\langle -0.4749;\ -0.1144\ \rangle$
Vomiting *	$\langle -0.2994;\ -0.0441\rangle$	$\langle -0.3000;\ -0.0484\ \rangle$
Shakes	$\langle -0.1585;\ 0.1060\rangle$	$\langle -0.1594;\ 0.1047\ \rangle$
Pruritus *	$\langle -0.3929;\ -0.0632\rangle$	$\langle -0.3932;\ -0.0667\ \rangle$
Hypotension	$\langle -0.0902;\ 0.0902\rangle$	$\langle -0.0955;\ 0.0955\rangle$

Adverse reaction	Wald
Nausea *	$\langle -0.4708\ ;\ -0.1292\ \rangle$
Vomiting *	$\langle -0.2607;\ -0.0393\ \rangle$
Shakes	$\langle -0.1309\ ;\ 0.0809\ \rangle$
Pruritus *	$\langle -0.3790\ ;\ -0.0710\ \rangle$
Hypotension	0

The methods provide different sets of confidence intervals. In the case of hypotension, the Wald interval degenerated to a single point 0, leading to the misleading conclusion that if no events occur in the trials, they never can (Fig. 6). The Wald interval does not provide sensible answers and is unsuitable for use with proportions close to 0. The other two confidence intervals conclude that the difference between the two proportions is not statistically significant (the confidence intervals contain 0). This indicates that both the groups are comparable (Fig. 4 and Fig. 5).

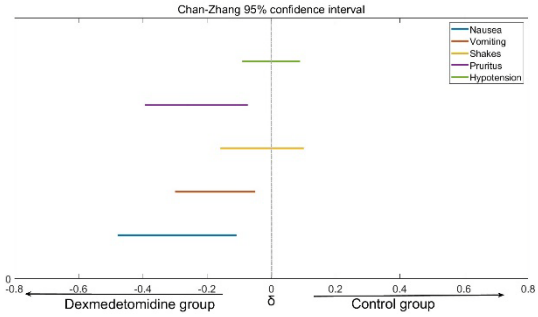

Figure 4. Chan-Zhang 95% confidence intervals for adverse reactions

For the remaining adverse reactions, Chan-Zhang, Agresti-Min and Wald intervals indicate that the incidence of nausea, vomiting and pruritus in the dexmedetomidine group was significantly lower than in the control group. The confidence intervals for these adverse reactions have negative endpoints and do not include 0. This indicates a higher proportion of patients with nausea, vomiting and pruritus in the control group (Fig. 4, Fig. 5 and Fig. 6).

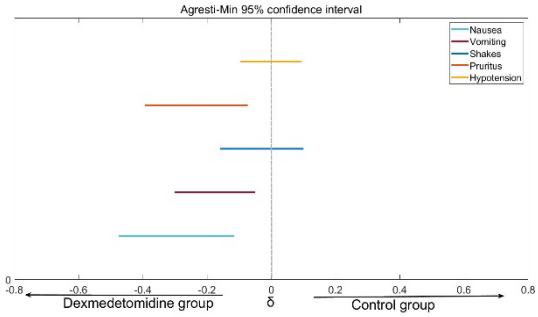

Figure 5. Agresti-Min 95% confidence intervals for adverse reactions

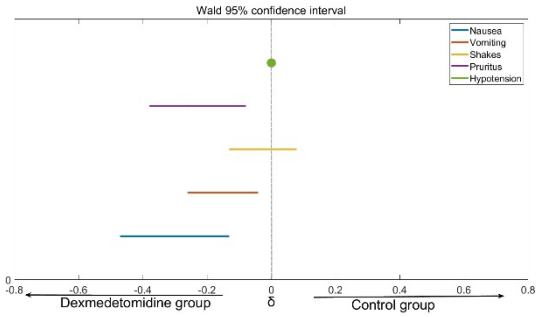

Figure 6. Wald 95% confidence intervals for adverse reactions

4. CONCLUSION

Confidence intervals are crucial in clinical trials, particularly for evaluating the efficacy and safety of new treatments against standards or placebos. The study by Mo and Qiu (2017) exemplifies the importance of this statistical tool, particularly when assessing adverse reactions to treatments. In their research, the exact Chan-Zhang and Agresti-Min intervals were used, both known for their strict conservatism and are particularly favoured in scenarios where conservatism is paramount.

However, the study also highlights the critical need for selecting appropriate interval calculation methods. The Wald interval, for example, was shown to be unsuitable near boundary values, such as a proportion close to 0, where it degenerated to a single point. This limitation can lead to misleading conclusions, especially in conditions with extremely low observed proportions.

Therefore, this study underscores the necessity of using more reliable methods like the Chan-Zhang and Agresti-Min intervals when dealing with proportions near boundaries. Such methods ensure accurate and meaningful statistical interpretations, enhancing the rigor and reliability of comparative medical research studies. The systematic use of well-chosen confidence intervals is essential for robust analysis outcomes and for avoiding potential pitfalls in medical research.

Sources

1. AGRESTI, A., CAFFO, B. *Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures*. 2000. American Statistician, Vol. 54, p. 280-288. ISSN 0003-1305.
2. AGRESTI, A., MIN, Y. *On small sample confidence intervals for parameters in discrete distributions*. 2001. Biometrics 57, p. 963-971. ISSN 0006-341X.
3. BROWN, D.L., LI, X. *Confidence intervals for two sample binomial distributions*. 2005. Journal of Statistical Planning and Inference 130, p. 359-375. ISSN 0378-3758.
4. CHAN, I.S.F., ZHANG, Z. Test-based exact confidence intervals for the difference of two binomial proportions. 1999. Biometrics 55, p. 1202-1209. ISSN 0006-341X.
5. GOMEZ-VAZQUEZ, M.E. et al. *Clinical analgesic efficacy and side effects of dexmedetomidine in the early postoperative period after arthroscopic knee surgery*. 2007. Journal of Clinical Anesthesia 19, p. 579-582. ISSN 0952-8180.
6. MIETTINEN, O. S., NURMINEN, M. Comparative analysis of two rates.1985. Statistics in Medicine 4, p. 213-226. ISSN:0277-6715.
7. MO, Y., QIU, S. *Effects of dexmedetomidine in reducing post-caesarean adverse reactions.* 2017. Experimental and Therapeutic Medicine, Vol. 14, p. 2036-2039. ISSN:1792-0981.
8. NEWCOMBE, R.G. *Interval estimate for the difference between independent proportions: comparison of eleven methods*. 1998. Statistics in Medicine, Vol. 17, p. 873-890. ISSN:0277-6715.
9. POBOČÍKOVÁ, I. *Exact and quasi-exact confidence intervals for the difference of two binomial proportions*. 2011. 10th International Conference Aplimat, Bratislava 2011, p. 1609-1617. ISBN 978-80-89313-51-8.