Approximate confidence intervals for a linear combination of binomial proportions: *A new variant* *

Sara Escudeiro, Adelaide Freitas, Vera Afreixo

S. Escudeiro, CIDMA, Polytechnic Institute of Coimbra, ESAC, 3040-316, Coimbra, Portugal
sarae@esac.pt
A. Freitas, CIDMA, Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
adelaide@ua.pt
V. Afreixo, CIDMA, IBIMED, Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
vera@ua.pt

Dated: —

Abstract

We propose a new adjustment for constructing an improved version of the Wald interval for linear combinations of binomial proportions, which addresses the presence of extremal samples. A comparative simulation study was carried out to investigate the performance of this new variant with respect to the exact coverage probability, expected interval length and mesial and distal non-coverage probabilities. Additionally, we discuss the application of a criterion for interpreting interval location in the case of small samples and/or in situations in which extremal observations exist. The confidence intervals obtained from the new variant performed better for some evaluation measures.

`keywrods:`Approximate confidence intervals Linear combination of binomial proportions Restricted models (Newcombe-Zou, Peskun and score methods) Unrestricted model (Wald method) Interval location

1 Introduction

Several approximate methods have been proposed in the literature for constructing confidence intervals (CIs) for one binomial proportion (?????), for the difference of two independent binomial proportions (???????) and, although in

*This is an Accepted Manuscript of an article published by Taylor & Francis in Communications in Statistics - Simulation and Computation on December/2017, available at http://www.tandfonline.com/doi/abs/10.1080/03610918.2016.1241408

smaller number, for any linear combination of $k \geq 2$ success proportions of independent binomial populations. In this last case, methods were proposed by **?** and **?** for $k = 2$ and by **?**, **?**, **?** and **?** for $k > 2$. The extension to $k > 2$ is particularly important, for instance, in the context of meta-analysis (**?**). The preference for asymptotic methods is generally justified because they are computationally faster and simpler to apply than exact ones.

Due to the dual relationship between statistical tests and CIs, the most common approach to obtain large-sample interval estimates for a combination θ of $k \geq 1$ binomial proportions $p_1, p_2, \ldots, p_k$ from independent binomial populations $X_1, X_2, \ldots, X_k$ with $n_1, n_2, \ldots, n_k$ trials, respectively, consists in inverting the standard two-sided Wald test $H_0 : \theta = \theta_0$. This test was proposed by **?** and is based on the normal approximation of maximum likelihood estimators (unrestricted model). The Wald test statistic is given by $\frac{(\theta - \hat{\theta})}{\sqrt{\hat{v}(\hat{\theta})}}$ and the general formula of the classic Wald CI is $\hat{\theta} \mp z_{\alpha/2}\sqrt{\hat{v}(\hat{\theta})}$, where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution and $\hat{\theta}$ and $\hat{v}(\hat{\theta})$ represent an estimate of θ and of the variance of estimator $\hat{\theta}$, respectively. For this type of CI, the variance is estimated using the maximum likelihood estimate (MLE) of θ, which depends on the MLEs of the k binomial proportions $p_1, p_2, \ldots, p_k$. A drawback of this CI are its poor coverage properties (e.g. **???**) and the occurrence of overshoot, because it is additively symmetric about the empirical estimate $\hat{\theta}$. The overshoot problem can be easily eliminated by truncating the support scale, but coverage may not be improved. Replacing the MLE of each p_i, $i = 1, 2, \ldots, k$, in the Wald test statistic, by a shrinkage estimator given by

$$\frac{X_i + h_i}{n_i + 2h_i}, \tag{1}$$

for some $h_i > 0$, that is, by adding h_i successes and h_i failures to the original data, the so-called adjusted Wald method is obtained. The adjusted Wald CIs are additively symmetrical about a mesially shrunk point estimate (**?**). Depending on the particular h_i chosen, different variants of the adjusted CIs can be established. When $h_i = 2/k$, we have the adjustment proposed by **?** for one proportion ($k = 1$), **?** for the difference between two proportions ($k = 2$) and **?** for a linear combination of proportions ($k > 2$). The adjusted Wald CIs have better performance than the classic Wald CI.

Another common approach, called the score method, is based on the score test proposed by **?**. In this method, the test statistic is subject to the constraint defined by the null hypothesis, $H_0 : \theta = \theta_0$, unlike the Wald method, which requires only the unrestricted (saturated) model in the estimation process. The bounds of the score CI are determined by solving the following equation in order to θ_0:

$$\left.\frac{\hat{\theta} - \theta_0}{\sqrt{v(\hat{\theta})}}\right|_{\theta=\theta_0} = z_{\alpha/2}\,. \tag{2}$$

Obviously, the complexity of solving (2) depends on the complexity of the estimator $\hat{\theta}$ and its variance. The score CI was first proposed for one proportion by **?**. A score-type CI for the difference of proportions was suggested by **?** where the parameters p_i on which the variance in (2) depends are substituted by their respective MLE's obtained under the assumption of $p_1 - p_2 = \theta_0$. A slightly

more conservative version of the Mee CI was constructed by **?** through the inclusion of a correction factor in variance $v(\hat{\theta})$, which substantially improved the coverage probability (**?**). **?** proposed a hybrid score CI for the difference between two proportions, where the limits of the score CIs, obtained separately for each proportion p_1 and p_2, were considered in determining the CI for $p_1 - p_2$. By using a similar procedure, CIs for any linear combination of $k \geq 2$ proportions were obtained by **?** and **?**. Although apparently first presented by **?**, this approach, which requires knowing, individually, the score CIs for each parameter whose coverage probabilities are close to the nominal level, was later named as the Method Of Variance Estimates Recovery (MOVER) in **?**. **?** also proposed a CI for the difference between two proportions, using the Lagrange's multipliers to minimize θ_0 subject to the constraint defined by equation (2). An extension of the Peskun procedure for a general linear combination of $k \geq 2$ binomial proportions was recently studied in **?**.

Adjusted versions of the score-type CIs, as those constructed from the adjusted Wald method by adding h_i successes and h_i failures to the n_i original observations from the binomial population X_i, and of the CIs resulting from the application, as suggested by **?**, of a continuity correction (cc) to the classical versions of the methods, were considered and empirically evaluated for the case of a single proportion (**?**) and the difference of two proportions (**?**). The results of their simulations showed that while some adjustment-types improve the performance of the classic Wald CI, the same does not hold for the score CI. Furthermore, the considered cc showed off, in general, to be useless. Since the midpoint of the score interval is located between the empirical estimate and the midpoint of the support scale, the addition of pseudo-frequencies in the score method would lead to intervals too distally located, i,e., too far out from the midpoint of the support scale. Accordingly, it is likely that, as a consequence, the score interval has poor coverage relative to length.

Being particularly focused on asymptotic CIs based on parameter estimation under the unrestricted model (Wald CIs) and restricted models (Newcombe-Zou, Peskun and score CIs), in the context of any linear combinations of two or more proportions, **?** proposed a new choice for h_i in (1) that establishes a generalization of the adjustment proposed by **?** when the minimum and maximum boundaries of successes (i.e., extreme observations) are reached. Considering the way in which the adjusted Wald CIs, which are computationally the least complex, are constructed, we now suggest an extension of the adjustment proposed by **?**, which takes into account the weights of the proportions in the linear combination. A simulation study based on the exact coverage probabilities and the expected interval length is carried out to compare the performance of this new adjusted version with that of the approximated CIs for a linear combination of proportions recently considered in the literature. Furthermore, we analyse the location of the CIs relatively to the centre of the support scale, based on the concept of the mesial and distal non-coverage probabilities introduced by **?** for one proportion. Although, interval location is a feature considered by relatively few researchers, it is useful for evaluating the suitability of the CIs methods analysed.

2 Motivating examples

We show two examples of application of the CIs for linear combinations.

Example 1 To illustrate the analysis of the effect of categorical factors, we use the data from the example described by **?**, where the effect of four different diets combining fibre and fat on the development of chemically induced tumours in rats is investigated. The population under study consists of rats divided in $k = 4$ groups. From each group, $n_i = 30\,(\forall i)$ rats were randomly selected. The four groups and the data from the experiment are reproduced in Table 1, where the success number x_i of rats with tumour was recorded.

Table 1: Diet and tumour study.

	Fibre		No Fibre	
	High fat	Low fat	High fat	Low fat
Sample size (n_i)	30	30	30	30
Rats with tumour (x_i)	20	14	27	19
Fibre × Fat	**+1**	**−1**	**−1**	**+1**
Fibre	+1	+1	−1	−1
Fat	+1	−1	+1	−1

If we denote by p_i the unknown population proportion in the i-th group, the existence of a linear-scale interaction between fibre and fat, for example, could be analysed by testing

$$H_0 : p_1 - p_2 - p_3 + p_4 = 0 \quad vs \quad H_1 : p_1 - p_2 - p_3 + p_4 \neq 0\,. \tag{3}$$

Due to the dual relationship between statistical tests and CIs, we can address the testing problem in terms of the CI for the linear combination $p_1 - p_2 - p_3 + p_4$. By inverting hypothesis test (3), the CI for the linear combination of the four independent binomial proportions can be constructed.

Example 2 In this example, we use the data from a meta-analysis of the diagnostic accuracy of non-contrast computed tomography on adults with suspected appendicitis, discussed in **?**. These data consists of seven studies where the specificity was perfect (100%) in four studies. The data are reproduced in Table 2, where the value of the specificity of each study was recorded. The aim of this example is to find an interval estimation for the pooled specificity, that is, for the linear combination $\sum_{i=1}^{7} \beta_i p_i$. For simplicity, we assume that all seven studies have the same weight $\beta_i = 1/7,\ \forall i$. It is convenient to refer that the evaluation of estimation processes in meta-analysis contexts is out of the scope of this paper.

Regarding the first example, in Section 5 we will construct all the approximate CI types mentioned in Section 1. In the second example, the analysis will be carried out using the Monte Carlo method, as the calculation of the exact coverage probabilities is a computationally intensive process. For both, we will discuss the best CI method.

Table 2: Non-contrast computed tomography for diagnosing appendicitis (**?**). TP: true positive; FP: false positive; FN: false negative; TN: true negative.

Study name	TP	FP	FN	TN	n_i $(TN+FP)$	Specificity $(TN/(TN+FP))$
Ashraf 2006	21	0	2	35	35	1.00
Ege 2002	104	3	4	185	188	0.98
Horton 2000	37	0	1	11	11	1.00
In't Hof 2004	83	0	4	16	16	1.00
Keyzer 2005	26	5	4	59	64	0.92
Stacher 1999	21	0	1	34	34	1.00
Tamburrini 2007	73	13	8	310	323	0.96

3 CIs for linear combination of proportions

Let $X_1, \ldots, X_k$ be $k \geq 2$ independent binomial random variables with parameters n_i and p_i, $i = 1, 2, \ldots, k$. Our interest lies in finding approximate CIs for a linear combination of binomial proportions defined as $L = \sum_{i=1}^{k} \beta_i \, p_i$, where $\beta_i \neq 0$ is a fixed value and p_i is the unknown population proportion. The MLE of L is equal to $\hat{L} = \sum_{i=1}^{k} \beta_i \, \hat{p}_i$, where $\hat{p}_i = \frac{X_i}{n_i}$ is the MLE of the proportion p_i. When the k proportions p_i are estimated using the shrinkage estimator (1), with $h_i \neq 0$, other estimates of the interesting parameter L, given by $\tilde{L} = \sum_{i=0}^{k} \beta_i \, \tilde{p}_i$, with $\tilde{p}_i = \dfrac{x_i + h_i}{\tilde{n}_i}$ and $\tilde{n}_i = n_i + 2h_i$, can be obtained. When suitably centred and scaled, the statistic $\hat{L}$ converges asymptotically in distribution to a standard normal distribution, i.e., $\frac{\hat{L}-L}{\sqrt{v(\hat{L})}} \xrightarrow{d} N(0,1)$, where $v(\hat{L})$ denotes the variance of $\hat{L}$ and is given by $v(\hat{L}) = \sum_{i=1}^{k} \frac{\beta_i^2 \, p_i \, (1-p_i)}{n_i}$. Using the unrestricted maximum likelihood model (Wald method) or restricted models (score method), approximate CIs for L can be constructed by inverting the two-sided test

$$H_0 : L = \lambda_0 \;\; vs \;\; H_1 : L \neq \lambda_0 \,, \tag{4}$$

where λ_0 is any real constant admissible for $\sum_{i=1}^{k} \beta_i \, p_i$, meaning that λ_0 should belong to the support scale $\left[\sum_{\beta_i<0} \beta_i \; ; \; \sum_{\beta_i>0} \beta_i\right]$. This inversion procedure leads to the Wald and score CIs given by

$$\left\{ \lambda_0 \in \left[\sum_{\beta_i<0} \beta_i \; ; \; \sum_{\beta_i>0} \beta_i \right] \; : \; |Z| < z_{\alpha/2} \right\}, \tag{5}$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution. The statistic Z depends on the chosen method, that is, the Wald test statistic (Z_W) for Wald CIs and the score test statistic (Z_S) for score CIs, which are defined as

$$Z_W = \frac{\hat{L} - \lambda_0}{\sqrt{\hat{v}(\hat{L})}} \quad \text{and} \quad Z_S = \frac{\hat{L} - \lambda_0}{\sqrt{\bar{v}(\hat{L})}} \,, \tag{6}$$

where $\hat{v}(\hat{L}) = \sum_{i=1}^{k} \frac{\beta_i^2 \, \hat{p}_i \, (1-\hat{p}_i)}{n_i}$ and $\bar{v}(\hat{L}) = \sum_{i=1}^{k} \frac{\beta_i^2 \, \bar{p}_{i0} (1-\bar{p}_{i0})}{n_i}$ represent, respectively, the estimates of variance $v(\hat{L})$ under the unrestricted model and the

restricted model ($\bar{p}_{i0}$ is the restricted MLE of p_i under the null hypothesis $H_0 : \sum_{i=1}^{k} \beta_i \, p_i = \lambda_0$). The expressions that define the lower and upper limits of the (classic) Wald and score CIs (5) for L are obtained by solving equation $Z^2 = z^2_{\alpha/2}$. In equation (6), if the proportions p_i of which variance $v(\hat{L})$ depends on are replaced by estimates based on (1) with $h_i \neq 0$, new adjusted versions, also known as variants, of the Wald and score methods can be deduced. A detailed deduction of the expressions that define the limits of the approximate CIs for L, obtained by applying either the classical and the adjusted versions of the Wald and score methods, can be found in **?**. Besides these two methods and their adjusted versions, other procedures for obtaining approximate CIs for linear combinations, such as the Peskun procedure and the MOVER approach in their classical and adjusted forms, have also been investigated in the literature, namely in **?** and **?**. The purpose of this paper is to deduce a new adjustment and to comparatively evaluate its performance against the other CIs herein mentioned. Before describing our proposal, we provide a brief synthesis of the formulas of those CIs. The following additional notation will be used: $\tilde{n} = \sum_{i=1}^{k} \tilde{n}_i$ and $B = \sum_{i=1}^{k} \beta_i$.

3.1 The Wald CIs

The lower and upper limits of the classic Wald CI are given by

$$\hat{L} \mp z_{\alpha/2} \sqrt{\hat{v}(\hat{L})} \, . \tag{7}$$

The expressions of the adjusted Wald CIs are obtained by replacing, in equation (7), $\hat{L}$ by $\tilde{L}$ and variance $\hat{v}(\hat{L})$ by the adjusted estimated variance given by $\tilde{v}(\hat{L}) = \sum_{i=1}^{k} \frac{\beta_i^2 \, \tilde{p}_i \, (1-\tilde{p}_i)}{\tilde{n}_i}$.

3.2 The score CIs

In the score method, the estimated variance $\bar{v}(\hat{L})$ under H_0 depends on λ_0, which satisfies the equation

$$n + (B - 2\lambda_0) \, \frac{Z_S^2}{\hat{L} - \lambda_0} - \sum_{i=1}^{k} R_i = 0 \, , \tag{8}$$

where $R_i^2 = \left(\beta_i Z_S^2 / \left(\hat{L} - \lambda_0 \right) + n_i \right)^2 - 4 \beta_i \, n_i \hat{p}_i Z_S^2 / \left(\hat{L} - \lambda_0 \right)$. A detailed deduction of equation (8) can be found in **?**. Making $Z_S^2 = z^2_{\alpha/2}$ and solving (8) in order to λ_0 by a numerical method, two solutions will be obtained: $\lambda_- < \lambda_+$. The extra condition $\sum_{\beta_i<0} \beta_i \leq \lambda_- < \hat{L} < \lambda_+ \leq \sum_{\beta_i>0} \beta_i$ is taken into account. Thus, the lower and upper limits of the classic score CI are given by

$$\max \left(\sum_{\beta_i<0} \beta_i, \lambda_- \right) \quad \text{and} \quad \min \left(\lambda_+, \sum_{\beta_i>0} \beta_i \right) \tag{9}$$

respectively.

For the adjusted score CIs, variance $v(\hat{L})$ will be estimated by $\tilde{v}(\hat{L}) = \sum_{i=1}^{k} \frac{\beta_i^2 \, \tilde{p}_{i0} \, (1-\tilde{p}_{i0})}{\tilde{n}_i}$,

where

$$\tilde{p}_{i0} = \frac{\tilde{n}_i + \beta_i \tilde{C} - \tilde{R}_i}{2\beta_i \tilde{C}}\,,\ \tilde{C} = \frac{Z_S^2}{\tilde{L} - \lambda_0} \quad \text{and} \quad \tilde{R}_i^2 = \left(\beta_i \tilde{C} + \tilde{n}_i\right)^2 - 4\beta_i\, \tilde{n}_i \tilde{p}_i \tilde{C}. \tag{10}$$

The corresponding adjusted version of equation (8) will therefore be

$$\tilde{n} + (B - 2\lambda_0)\, \frac{Z_S^2}{\tilde{L} - \lambda_0} - \sum_{i=1}^{k} \tilde{R}_i = 0.$$

Analogously, two solutions are found by using a numerical method: $\tilde{\lambda}_- < \tilde{\lambda}_+$. Consequently, conditioned by $\sum_{\beta_i<0} \beta_i \leq \tilde{L} \leq \sum_{\beta_i>0} \beta_i$, the lower and upper limits of the adjusted score CIs are given by

$$\max\left(\sum_{\beta_i<0} \beta_i, \tilde{\lambda}_-\right) \quad \text{and} \quad \min\left(\tilde{\lambda}_+, \sum_{\beta_i>0} \beta_i\right) \tag{11}$$

3.3 The Newcombe-Zou CIs

The construction of the classic Newcombe-Zou CI for L is based on the MOVER approach, which starts by computing the Wilson CI for each single proportion p_i given by the following lower (a_{i_-}) and upper (a_{i_+}) bounds:

$$a_{i_\mp} = \frac{\hat{p}_i\, n_i + \dfrac{z_{\alpha/2}^2}{2} \mp z_{\alpha/2} \sqrt{\dfrac{z_{\alpha/2}^2}{4} + \hat{p}_i\,(1 - \hat{p}_i)\, n_i}}{n_i + z_{\alpha/2}^2}\,. \tag{12}$$

The lower and upper limits of the classic Newcombe-Zou CI are given by

$$\hat{L} \mp z_{\alpha/2} \sqrt{\sum_{\beta_i<0} \frac{\beta_i^2\, a_{i_\pm}\,(1 - a_{i_\pm})}{n_i} + \sum_{\beta_i>0} \frac{\beta_i^2\, a_{i_\mp}\,(1 - a_{i_\mp})}{n_i}}\,. \tag{13}$$

The expressions of the adjusted Newcombe-Zou CIs are obtained by replacing, in equation (13), $\hat{L}$ by $\tilde{L}$ and the variance by the adjusted estimated variance given by $\sum_{\beta_i<0} \frac{\beta_i^2\, \tilde{a}_{i_\pm}\,(1 - \tilde{a}_{i_\pm})}{\tilde{n}_i} + \sum_{\beta_i>0} \frac{\beta_i^2\, \tilde{a}_{i_\mp}\,(1 - \tilde{a}_{i_\mp})}{\tilde{n}_i}$, where

$$\tilde{a}_{i_\mp} = \frac{\tilde{p}_i\, \tilde{n}_i + \dfrac{z_{\alpha/2}^2}{2} \mp z_{\alpha/2} \sqrt{\dfrac{z_{\alpha/2}^2}{4} + \tilde{p}_i\,(1 - \tilde{p}_i)\, \tilde{n}_i}}{\tilde{n}_i + z_{\alpha/2}^2}\,. \tag{14}$$

3.4 The Peskun CIs

The classic Peskun CI for a linear combination L is based on a general method for obtaining confidence limits from a sample from the sampling distribution of the MLE $\hat{L}$. Taking into account the normal limiting distribution of $\hat{L}$ and given an observed value l_0 of $\hat{L}$, the method determines the minimum and maximum values of the function $L = \sum_{i=1}^{k} \beta_i\, p_i$ as being the lower and upper $100(1 - \alpha)\%$-confidence limits for L, respectively, subject to the condition

$(l_0 - L)^2 / v(\hat{L}) = z^2_{\alpha/2}$. This constraint is equivalent to $P(|\hat{L}| < l_0) \approx 1 - \alpha$, which ensures the nominal level $(1-\alpha)$ for the CI in construction. The method of Lagrange multipliers will be applied to find the solution for each of the two above mentioned optimization problems. The classic Peskun CI is given by

$$\frac{n}{n + z^2_{\alpha/2}} \left\{ \hat{L} + \frac{B z^2_{\alpha/2}}{2n} \mp \frac{z_{\alpha/2}}{2} \sqrt{\frac{n + z^2_{\alpha/2}}{n} \left(\sum_{i=1}^{k} \frac{\beta_i^2}{n_i} \right) - \frac{\left(B - 2\hat{L} \right)^2}{n}} \right\}. \quad (15)$$

The expressions of the adjusted Peskun CIs are obtained by replacing $(\hat{L}, n_i, n)$ in (15) by $(\tilde{L}, \tilde{n}_i, \tilde{n})$.

4 A new adjustment

Roughly speaking, the expressions of the boundaries of all CIs described above depend on the estimator $\tilde{p}_i$ of p_i, which is defined in terms of the parameter h_i. Thus, each new choice for h_i establishes a new variant of the method for constructing approximate CIs for L. When $h_i = 0$, we obtain the MLE of p_i ($\tilde{p}_i = \hat{p}_i$) and, therefore, the classical version of the method. On the other hand, if $h_i > 0$, adjusted versions of the method are obtained. By setting $h_i = 1$, we obtain the Laplace estimator suggested by **?**, which was also considered by **?** for the interval estimation of linear combinations of $k > 2$ binomial proportions. If $h_i = 2/k$, we obtain the estimator proposed by **?** for one proportion ($k = 1$ and $\beta_1 = 1$), the estimator proposed by **?** for the difference between two proportions ($k = 2$, $\beta_1 = 1$ and $\beta_2 = -1$) and the estimator introduced by **?** for any linear combination of proportions ($k > 2$).
There is no optimal solution h_i for which $\tilde{p}_i$ has the smallest mean square error (MSE) for all p_i. Considering the average MSE with respect to a uniform prior distribution on $[0, 1]$, the Bayes risk is minimized when $h_i = 1$ (**?**). Unlike $\hat{p}_i$, the estimator $\tilde{p}_i$ is biased when $h_i > 0$ and unbiased at the midpoint of the support scale $p_i = 1/2$, for all $h_i \geq 0$, with $\text{MSE}(\tilde{p}_i) = n_i/(4(n_i + 2h_i)^2)$. In these circumstances, $\tilde{p}_i$ can then be interpreted as a shrinkage estimator of p_i for which the degree of shrinkage depends on n_i. For some chosen h_i, the degree of shrinkage towards the midpoint $1/2$ is high when n_i is small and low for large n_i (**?**). Since the classic Wald CI is based on $\hat{p}_i$ and has its worst performance when $x_i \in \{0, n_i\}$, one way to push the estimate closer to $1/2$ would be to consider the estimator (1) with a higher value of $h_i > 0$, when extremal samples occur. This is what **?** intended to achieve with the new adjustment h_i they proposed for CIs of $k \geq 2$ proportions. Their adjustment is based on the same type of reasoning that Agresti and Coull followed for one proportion (both the Agresti and Coull interval and the Wilson interval are centred at the same value). Concretely, those authors established a new estimator belonging to (1) by setting a value for h_i that leads the centre of the adjusted Wald CI (based on the unrestricted model) to be approximated to the centre of the classic score CI (based on a restricted model). Exact formulas for the centre of the classic score CI are not known but we can find approximations for them (more details in Appendix A).

One approximation is given by

$$\hat{L} + \frac{B^* \, z^2_{\alpha/2}}{2\left(N^* + z^2_{\alpha/2}\right)} + \frac{z^2_{\alpha/2}}{2} \, f, \tag{16}$$

where

$$f = \frac{\sum_{i=1}^{k} \dfrac{\beta_i^3 \hat{p}_i (1-\hat{p}_i) b_i}{n_i^2}}{\displaystyle\sum_{i=1}^{k} \frac{\beta_i^2 \, \hat{p}_i \, (1-\hat{p}_i)}{n_i}}, \quad N^* = \sum_{i=1}^{k} n_i \mathbf{1}_{\mathcal{A}_i}(x_i), \quad B^* = \sum_{i=1}^{k} b_i \beta_i \mathbf{1}_{\mathcal{A}_i}(x_i), \quad b_i = 1-2\hat{p}_i,$$

with $\mathbf{1}_{\mathcal{A}_i}(x)$ being the indicator function of the set

$$\mathcal{A}_i = \left\{ x_i \in \{0, n_i\} : \; (n_i - 2x_i)\,(\hat{L} - \lambda_0)\,\beta_i < 0 \right\}. \tag{17}$$

Remark that each set $\mathcal{A}_i$ can only contain one observation, which will be an extreme observation ($x_i = 0$ or $x_i = n_i$) and will depend on whether we are computing the lower limit ($\lambda_0 < \hat{L}$) or the upper limit ($\lambda_0 > \hat{L}$) of the CI (see more details in Appendix A). Since the centre of the adjusted Wald CI is given by $\sum_{i=1}^{k} \beta_i \tilde{p}_i$, then, imposing the score and adjusted Wald CIs to have the same centre, we have

$$\sum_{i=1}^{k} \beta_i \frac{x_i + h_i}{n_i + 2h_i} = \sum_{i=1}^{k} \beta_i \frac{x_i}{n_i} + \sum_{i=1}^{k} \beta_i \frac{b_i \, z^2_{\alpha/2} \mathbf{1}_{\mathcal{A}_i}(x_i)}{2\left(N^* + z^2_{\alpha/2}\right)} + \frac{z^2_{\alpha/2}}{2} \, f, \tag{18}$$

Thus, the following term-by-term equality emerges:

$$\frac{x_i + h_i}{n_i + 2h_i} = \frac{x_i}{n_i} + \frac{b_i \, z^2_{\alpha/2} \mathbf{1}_{\mathcal{A}_i}(x_i)}{2\left(N^* + z^2_{\alpha/2}\right)} + \frac{z^2_{\alpha/2}}{2} \, \frac{\dfrac{\beta_i^2 \hat{p}_i (1-\hat{p}_i) b_i}{n_i^2}}{\displaystyle\sum_{i=1}^{k} \frac{\beta_i^2 \, \hat{p}_i \, (1-\hat{p}_i)}{n_i}}. \tag{19}$$

? suggested to approximate f in (18) by assuming each i-th term $\dfrac{\beta_i^2 \hat{p}_i (1-\hat{p}_i)}{n_i}$ to be the arithmetic average of all i-th terms, that is,

$$\frac{\beta_i^2 \hat{p}_i (1-\hat{p}_i)}{n_i} \simeq \frac{1}{k} \sum_{i=1}^{k} \frac{\beta_i^2 \, \hat{p}_i \, (1-\hat{p}_i)}{n_i}. \tag{20}$$

After substituting expression (20) into (19), those authors obtained

$$h_i \simeq \frac{z^2_{\alpha/2}}{2} \left(\mathbf{1}_{\mathcal{A}_i}(x_i) + \frac{1}{k} \right). \tag{21}$$

When there are no extremal observations, the approximation (21) becomes simpler: $h_i \simeq z^2_{\alpha/2}/2k$. Remark that $h_i = z^2_{\alpha/2}/2k$ corresponds to the choice of h_i proposed by **?** for the parametric family (1).

Choice (21) does not establish any dependence of the estimate found for p_i on the linear combination L, which is the primary parameter in the estimation process. Intuitively, the higher the β_i the more influence the estimate of p_i is expected to have on the estimation of L and the more impact there will be on the variance of $\hat{L}$ the in terms of β_i^2/n_i. In particular, for unbalanced experimental designs and taking into account the term β_i^2/n_i, the empirical proportions of success associated to the higher weights on the linear combination should expectedly have a higher effect on the estimation of L when the corresponding sample sizes are smaller. We suggest considering an alternative approximation for f by taking all the terms $\hat{p}_i(1-\hat{p}_i)$ as constant, that is,

$$\hat{p}_i(1-\hat{p}_i) \simeq \frac{\displaystyle\sum_{i=1}^{k} \frac{\beta_i^2\, \hat{p}_i\,(1-\hat{p}_i)}{n_i}}{\sum_{i=1}^{k} \dfrac{\beta_i^2}{n_i}} \,. \tag{22}$$

Under this assumption, if we substitute expression (22) into (19), we obtain

$$h_i \;\simeq\; \frac{\dfrac{z_{\alpha/2}^2}{2}\left(\dfrac{n_i \mathbf{1}_{\mathcal{A}_i}(x_i)}{N^* + z_{\alpha/2}^2} + \dfrac{\beta_i^2}{n_i \sum_{i=1}^{k} \dfrac{\beta_i^2}{n_i}}\right)}{1 - \dfrac{n_i \mathbf{1}_{\mathcal{A}_i}(x_i)}{N^* + z_{\alpha/2}^2}\dfrac{z_{\alpha/2}^2}{n_i} - \dfrac{z_{\alpha/2}^2 \beta_i^2}{n_i^2 \sum_{i=1}^{k} \dfrac{\beta_i^2}{n_i}}} \tag{23}$$

For a large enough n_i, the expression $\dfrac{n_i}{N^* + z_{\alpha/2}^2} = \dfrac{1}{1+\mathrm{O}(1/n_i)}$ could be approximated to one and hence, in view of (23), we suggest taking

$$h_i \simeq \frac{z_{\alpha/2}^2}{2}\left(\mathbf{1}_{\mathcal{A}_i}(x_i) + \frac{\beta_i^2/n_i}{\sum_{i=1}^{k} \dfrac{\beta_i^2}{n_i}}\right), \tag{24}$$

which makes estimate (1) depend on (β_i, n_i), in terms of β_i^2/n_i, as we intended to.

Table 3 summarizes some of the choices for parameter h_i established in the literature and also includes the new choice (24) we propose in this paper. As mentioned previously, each choice for h_i establishes a different variant of the method for constructing approximate CIs for L. The variants corresponding to each of the choices for h_i in Table 3 are herein designated sequentially by number, from variant-0 to variant-4. Variant-4 is the result of the new h_i we propose. The four adjustments can be applied to any of the four CI methods (Wald, score, Peskun and Newcombe-Zou) described in the previous section. It is obvious that variant-2 is equal to variant-3 when $0 < x_i < n_i$ for all i, variant-1 is approximately equal to variant-2 when $\alpha = 5\%$, and variant-3 is equal to variant-4 when $\dfrac{\beta_i^2}{n_i}$ is a constant for all i. We call attention to the fact that parameter h_i in both variant-3 and variant-4 is different when we are computing the lower

limit ($\mathcal{A}_i$ will be given by $\{x_i : (x_i = 0 \wedge \beta_i < 0) \vee (x_i = n_i \wedge \beta_i > 0)\}$) or the upper limit ($\mathcal{A}_i$ will be given by $\{x_i : (x_i = 0 \wedge \beta_i > 0) \vee (x_i = n_i \wedge \beta_i < 0)\}$) of the CIs. Thus, the value of h_i in variant-3 and variant-4 increases in the presence of the extremal samples $x_i \in \{0, n_i\}$. Therefore, when the true proportion coincides with the midpoint of the support scale, $p_i = 1/2$, the estimator $\tilde{p}_i$ in variant-3 and variant-4 will push estimates closer to the true proportion, especially whenever extremal samples occur. Concerning the midpoint or the case $h_i = z^2_{\alpha/2}/2k$, the centre (16) of the score CI will be $\hat{L} + f\, z^2_{\alpha/2}$, which is an approximation of the centre of the 95%-confidence CI of variant-1 of the Wald method when f is approximated by 1, and of the $100(1-\alpha)$%-confidence CI of variant-2 of the Wald method when f is approximated by $1/k$.

From the expressions of h_i for variants-3, 4, it is obvious that both of them are affected by the presence of extremal observations. In order to evaluate the similarity between the estimates of the proportions p_i when variants-3, 4 of the adjusted Wald method are considered for the construction of the CI for L, we suggest calculating the euclidean distance

$$d = \sqrt{\sum_{i=1}^{k} \left(\frac{\beta_i^2}{n_i \sum_{i=1}^{k} \frac{\beta_i^2}{n_i}} - \frac{1}{k} \right)^2}. \tag{25}$$

Table 3: Parameters h_i of the classic and adjusted variants, for the estimation of p_i based on the shrinkage estimator (1).

	classic	adjusted			
	variant-0	variant-1	variant-2	variant-3	variant-4
h_i	0	$\frac{2}{k}$	$\frac{z^2_{\alpha/2}}{2k}$	$\frac{z^2_{\alpha/2}}{2}\left(\mathbf{1}_{\mathcal{A}_i}(x_i) + \frac{1}{k}\right)$	$\frac{z^2_{\alpha/2}}{2}\left(\mathbf{1}_{\mathcal{A}_i}(x_i) + \frac{\beta_i^2/n_i}{\sum_{i=1}^{k} \frac{\beta_i^2}{n_i}}\right)$

If β_i^2/n_i is a constant for all i then $d = 0$, which in turn indicates that the two variants coincide and so they will produce the same results. If $d > 0$, contributions of the weights β_i, in terms of β_i^2/n_i, may be expected in the estimation of the singular proportions p_i of the linear combination L. This means that the higher the value of d, the bigger the contribution of these weights and the lower the similarity of the results provided by variants-3, 4.

5 Evaluation of the CI methods

5.1 Evaluation measures

In order to assess and compare the performance of the twenty (four procedures $\times$ five adjustments) approximate CIs discussed in the previous sections, we present an evaluation of their exact coverage probabilities, expected lengths and locations via simulation. The location of a CI is characterized by its mesial and

distal non-coverage probabilities. This evaluation was performed for the cases of $k = 3$ and $k = 4$ independent binomial populations. Three confidence levels were analysed: 90%, 95% and 99%. Since similar conclusions were obtained in the two cases for the three nominal levels, only the results for 95% will be herein discussed. All simulations were carried out using R scripts, with the numerically intensive calculations implemented as C functions. The source code (in R and C) is available on request from the first author.

Given the parameters (n_i, β_i) and a set of k binomial proportions $(p_1, p_2, \ldots, p_k)$, the exact coverage probability (R) and the expected interval length (L) are defined as

$$\mathrm{R} = \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \cdots \sum_{x_k=0}^{n_k} \prod_{i=1}^{k} \binom{n_i}{x_i} p_i^{x_i} (1-p_i)^{n_i - x_i} \, \mathbf{1}_{[l(\boldsymbol{x}), u(\boldsymbol{x})]}(L)$$

and

$$\mathrm{L} = \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \cdots \sum_{x_k=0}^{n_k} \prod_{i=1}^{k} \binom{n_i}{x_i} p_i^{x_i} (1-p_i)^{n_i - x_i} \, \left(u\left(\boldsymbol{x}\right) - l\left(\boldsymbol{x}\right) \right),$$

where $[l\left(\boldsymbol{x}\right), u\left(\boldsymbol{x}\right)]$ is the CI obtained from the observation vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ for the linear combination $L = \sum_{i=1}^{k} \beta_i \, p_i$. To examine the interval location, we extend the procedure suggested by **?** for the case of one proportion. Concretely, for each CI of a linear combination, we analyse the existence of equilibrium between the directions of the mesial non-coverage probability (MNR) and distal non-coverage probability (DNR), in the form MNR=DNR. These directions indicate whether the CIs are located too distally or too mesially from the midpoint of the support scale $c = \sum_{i=1}^{k} \beta_i / 2$ relatively to the true value L of the linear combination. The MNR and DNR are defined as

$$\mathrm{MNR} = \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \cdots \sum_{x_k=0}^{n_k} \prod_{i=1}^{k} \binom{n_i}{x_i} p_i^{x_i} (1-p_i)^{n_i - x_i} \, \mathbf{1}_{\mathcal{M}}(\boldsymbol{x}),$$

with $\mathcal{M} = \{\boldsymbol{x} : (L \leq c \wedge u\left(\boldsymbol{x}\right) < L) \vee (L \geq c \wedge l\left(\boldsymbol{x}\right) > L)\}$, and

$$\mathrm{DNR} = \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \cdots \sum_{x_k=0}^{n_k} \prod_{i=1}^{k} \binom{n_i}{x_i} p_i^{x_i} (1-p_i)^{n_i - x_i} \, \mathbf{1}_{\mathcal{D}}(\boldsymbol{x}),$$

with $\mathcal{D} = \{\boldsymbol{x} : (L < c \wedge l\left(\boldsymbol{x}\right) > L) \vee (L > c \wedge u\left(\boldsymbol{x}\right) < L)\}$.

In our simulation, 10000 sets of k binomial proportions $(p_1, \ldots, p_k)$, with each p_i randomly generated from the standard uniform distribution, were used for each k, and the four quantities R, L, MNR and DNR were computed for each set of binomial proportions. Besides the calculation of the average values of R, L, MNR and DNR ($\mathrm{R_{mean}}$, $\mathrm{L_{mean}}$, $\mathrm{MNR_{mean}}$, $\mathrm{DNR_{mean}}$) for each linear coefficient and sample-size configuration and for each CI-type, two additional statistical measures recommended by **?** were calculated over the 10000 values of R: the minimum exact coverage probability estimates, $\mathrm{R_{min}}$, and the percentage of coverage probabilities that are lower than $100(1 - \alpha - 0.02)\%$, which is $\mathrm{R_{93\%}}$ for $\alpha = 5\%$. Remark that an interval is conservative if $\mathrm{R_{mean}} > 100(1-\alpha)\%$, whereas it is liberal if $\mathrm{R_{mean}} < 100(1-\alpha)\%$.

5.2 Evaluation rules

Based on the exact coverage probabilities and the expected interval length, **?** suggested three main rules applied in the following order for selecting the best method: (I) The method must have few liberal failures, i.e. the values of $R_{93\%}$ must be as small as possible; (II) The values of R_{mean} must be close to the nominal level of $100(1-\alpha)\%$; (III) The values of L_{mean} must be as small as possible.
We added a fourth rule based on the expected MNR and DNR. The interval location of the CIs can be characterized (**?**) by the ratio $Q = MNR/(1-R)$. This ratio expresses the balance condition between MNR and DNR. For a 95%-confidence level, the CIs for a linear combination are expected to produce $1 - R = MNR + DNR = 0.05 \quad \text{and} \quad MNR = DNR = 0.025$.

Based on a partition of the range of values of Q, **?** established a classification criterion for the interval location when $k = 1$ (one proportion) whose application we extend herein to the location of CIs for any linear combination ($k \geq 1$). Concretely, values of Q around 0.5 (between 0.375 and 0.625) are interpreted as corresponding to satisfactorily located interval estimates, less than 0.375 as intervals located too mesially to include the true value of L, and greater than 0.625 as intervals located too distally to include L. This classification assumes that the mesial and distal non-coverage probabilities are balanced, which is often not possible for CIs constructed from small sample sizes or extremal observations. In these situations, it seems more adequate to evaluate the interval location in terms of both the MNR and DNR, individually. The level of proximity of these two probabilities to the reference value 0.025 can be classified in the following way: values of DNR (MNR, resp.) between 0.02 and 0.03 will correspond to CI methods which yield intervals with a satisfactory mesial (distal) location. The lower limit of this interval is obtained by rounding $(1-R)\times 0.375 = 0.01875$ and the upper limit by rounding $(1-R)\times 0.625 = 0.03125$. Values of DNR (MNR, resp.) outside that range will correspond to non-satisfactory mesially (distally) located intervals. Selection of the best CI based on the individual values of the MNR and DNR is particularly relevant in contexts where mesially or distally shifted estimates are preferred. For instance, when for all or almost all k populations the values of the true proportions p_i are known beforehand to be close to the same boundary of the support scale $(0;1)$ of p_i, the interval estimation of any convex linear combination of these p_i will be better for CI methods which tend to err on the side of being mesially satisfactory (i.e., the DNR is close to 0.025 for 95% confidence intervals). In this context, we can refer to Example 2 as an example of pooled specificity estimation over $k = 7$ independent populations. Since the observed specificities of the seven studies are very high (four of them reach the maximum value), it seems plausible to suppose that the real specificity of the diagnostic test under evaluation is also very high. If practical expectations regarding the diagnostic test under evaluation suggest the occurrence of extreme observations, we will then prefer to select a CI method for the pooled specificity that besides being able to handle extremal observations it also has the tendency to err on the side of under- rather than overestimate the global specificity of the diagnostic test and, consequently, to ensure that the error of the mesially shifted point estimates is within the expected limits. The analysis of the performance of the interval estimation of linear combinations associated

to contexts where mesially or distally shifted estimates are arguably preferred will be interesting (e.g., very high proportions), but it is out of the scope of this paper.

Therefore, as follows from discuss above, we suggest adding a new rule for selecting the best method: (IV) The values of Q_{mean} must be between 0.375 and 0.625 and, for CIs where extremal observations are tolerated, the values of DNR_{mean} must be between 0.01875 and 0.03125. In contexts where mesially shifted estimates are preferred, this rule should be a selection criterion as important as the second rule and may establish an *a posteriori* reasonability criterion, as argued by **?**.

5.3 Simulation results

Our simulation study includes not only the parameter settings considered by **?**, but also four additional β_i-configurations (two for $k = 3$, the last two cases in Tables 4-6, and two for $k = 4$, the last two cases in Tables 8-10) that provide a greater variety of distance values d, defined in (25), and include settings with higher weights β_i for smaller sample sizes n_i (unbalanced scenarios, as mentioned in Section 4). Furthermore, we also add an assessment of the interval location for all settings considered. For $k = 3$, Tables 4 and 5 show the results of the evaluation measures described in Section 5.1 for the classic and the four adjusted Wald CIs, while Table 6 shows the results for the classical version of the CIs that are based on the score method. Due to the high percentage of liberal failures obtained (e.g.,$R_{93\%}$(variant-1)$\geq R_{93\%}$(variant-0) for the Newcombe-Zou, Peskun and score CIs) for all settings considered in the simulation study, the corresponding values of Table 6 for the adjusted versions of those CIs are not shown. Moreover, since the corresponding results for $k = 4$ are similar, they are included in Apenddix B for a 95% level of confidence (Table 8-10). Our reading of Tables 4-6 and 8-10 may be summarized as follows. It is important to remark that for the settings and evaluation measures discussed by **?**, our findings coincide, as expected, with those observed by those authors.

Wald CIs: variant-0, variant-1, variant-2, variant-3, variant-4

- Variant-0 is very liberal ($R_{mean} \leq 94.0\%$). The values of $R_{93\%}$ are too high, even for large sample sizes. It has the best L_{mean} values, because the R_{mean} values are below the nominal level. The interval location tends to be either satisfactory or too distal;

- Variant-1 and variant-2 are similar in trend because $z_{\alpha/2} \simeq 2$. Both are slightly conservative or slightly liberal. The performance of these two variants shows a tendency to improve when all the independent samples are large. When there are small sample sizes, the values of $R_{93\%}$ can become too high, being worse for variant-2. Although variant-1 has led to CIs that are slightly more mesially located than those from variant-2 (Q_{mean}(variant-1)$<Q_{mean}$(variant-2)), most interval locations seem satisfactory in both variants;

- Variant-3 and variant-4 are very conservative and their R_{mean} values are, in general, very similar. As expected, the results are the same in all

cases where β_i^2/n_i is a constant for all i ($d = 0$). Although the values of R_{min} for these two variants are closer to the nominal level 95% than those of the other variants, variant-4 provides a generally more satisfactory proportion ($R_{93\%}$ $\leq$1.3% for all cases considered) of liberal failures than variant-3. Since variants-3, 4 can handle extremal observations, the evaluation of the interval location for these two variants in terms of both the MNR and DNR is also included. The Q_{mean} values show, for most of the settings considered, an absence of equilibrium between the mesial and distal non-coverage probabilities in both variants, with a clear predominance of too mesially located CIs. In particular, we can conclude that (i) the proportion of a CI to be mesially located is satisfactory (DNR_{mean} is close to 2.5%), with variant-4 having the most stable values of DNR_{mean} across all settings, and (ii) the proportion of a CI to be distally located is very low for almost all settings, which is consistent with the capability of these two variants to handle extremal observations. For larger sample sizes, these two conclusions hold for variant-4 but not for variant-3, for which the values of Q_{mean} show that the interval location can become satisfactory or slightly mesial.

Newcombe-Zou, Peskun and score CIs: variant-0

- The Newcombe-Zou CI method is slightly conservative. Among the three classic CI methods, the Newcombe-Zou strategy is the one that provides CIs with the smallest range. The values of $R_{93\%}$ are usually high, particularly when compared to those from the score and Peskun CIs. For small sample sizes, these values are too high. The interval location is too mesial for all settings;

- The Peskun CI method is too conservative (R_{mean} >97.3%). The interval location is satisfactory or slightly too mesial in most of the settings. This is the method with the highest L_{mean} values. Therefore, despite the fact that, in most cases, its $R_{93\%}$ values are almost all zero, this method is not advisable;

- The score CI method is slightly liberal in some cases and slightly conservative in other cases (R_{mean} in $93.8\% - 95.6\%$). The L_{mean} values are very similar to those of the classic Wald CI. The values of $R_{93\%}$ never fail when all samples are large. The interval location almost always tends to be too mesial, or satisfactory when considering large sample sizes.

Although it is possible to say, from our simulation results, that the classic Wald method and all the adjusted score methods have been the worst performers, it is impossible to determine which is the best single method across all β_i-configurations. An analysis of the coverage probabilities shows that the classic Newcombe-Zou CI method yields the best global results ($95.2\% \leq R_{mean} \leq 95.4\%$). Based on the $R_{93\%}$ values, it is the classic Peskun CI method and variant-4 of the Wald method that show greater tendency to yield the lowest percentages of liberal failures and were, together with variant-3, the best performers in terms of the evaluation measure R_{min}, producing the values closest to the nominal level. Concerning the expected length, the highest accuracy CIs were produced by the classic Wald, Newcombe-Zou and score CI methods. The

best results in terms of Q_{mean} were produced by variant-1 and variant-2 of the Wald method, which produced satisfactorily located CIs. When considering the evaluation measures MNR_{mean} and DNR_{mean}, it is apparent that the CIs produced by variant-3 and variant-4 of the Wald method have satisfactory mesial locations and a very low probability of being distally located, which is consistent with the capability of these two variants to handle extremal observations.

Notwithstanding, based on the four ordered rules mentioned in Section 5.2, we can claim that the selection made by **?** when using the first three of those rules (namely, the score CIs as the best procedure, followed by variant-3 of the Wald CI as a much simpler alternative and having an acceptable good performance in terms of liberal failures and coverage probability, and finally, the Peskun CI, although it is too conservative and leads to excessively wide CIs) still holds, but with variant-4 as a better alternative than variant-3, particularly in more unbalanced scenarios, where there are fewer failures. Remark that both variant-3 and variant-4 yield intervals with mesially satisfactory locations. Although the Peskun procedure yields CIs with better location than the score procedure and almost never fails, it is also our third choice due to the two already mentioned drawbacks: it is too conservative and leads to excessively wide CIs.

Correction for continuity

Analogously to **?**, we also complement our evaluation by carrying out an analysis of the performance of the four procedures (Wald, Newcombe-Zou,Peskun and score) when the usual continuity correction (cc) proposed by **?** is applied (detailed information on how to define the limits of CIs with cc can be found in **?**). For $k = 4$, we found no such analysis in the literature. Our analysis showed that with the exception of the percentage of failures, the results of the evaluation measures are similar between the corrected and uncorrected versions of the methods for the twelve parameter settings considered (comparing the results presented in Tables 4-6 and Tables 8-10, the differences are smaller than 0.15). For $k = 3$ and $k = 4$, the percentage of failures produced by the corrected methods is always less than or equal to that of the uncorrected methods. In the case $k = 3$, there is a slight decrease of less than 1.1 units (in **?**, those differences were around 0.5 units) in the percentage of failures of the Newcombe-Zou and score CIs, both with cc, especially for small samples. However, for $k = 4$ this reduction practically ceases to exist, being less than 0.1 units (for more details, see Table 11 in Appendix B).

Table 4: Results of some of the evaluation measures for variants-0, 1, 2 of the Wald CI, for $k = 3$. Confidence level $1 - \alpha = 95\%$.

Method: **Wald**-95%	classic					adjusted									
$(\beta_1, \beta_2, \beta_3)$	variant-0					variant-1					variant-2				
$n_1/n_2/n_3$	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}
	(%)	(%)	(%)			(%)	(%)	(%)			(%)	(%)	(%)		
$(1/3, 1/3, 1/3)$															
10/10/10	91.6	28.2	89.7	0.27	0.617	95.6	91.8	0.2	0.28	0.322	95.5	91.8	0.2	0.28	0.332
30/30/30	94.0	63.0	3.7	0.16	0.579	95.2	93.1	0.0	0.16	0.406	95.2	93.1	0.0	0.16	0.413
30/10/10	91.6	47.8	96.0	0.24	0.613	95.5	91.9	0.1	0.25	0.358	95.4	91.9	0.1	0.25	0.367
30/20/10	92.2	52.0	83.1	0.22	0.609	95.4	89.9	0.1	0.22	0.379	95.3	89.9	0.1	0.22	0.387
$(1, 1, -1)$															
10/10/10	91.6	28.2	89.2	0.82	0.619	95.6	90.8	0.1	0.83	0.319	95.5	90.8	0.1	0.83	0.330
30/30/30	94.0	63.0	3.6	0.49	0.580	95.2	93.4	0.0	0.49	0.405	95.2	93.4	0.0	0.49	0.412
30/10/10	91.6	46.9	96.6	0.72	0.621	95.5	91.1	0.0	0.74	0.354	95.4	91.1	0.1	0.74	0.364
30/20/10	92.1	51.4	83.9	0.65	0.615	95.4	91.4	0.1	0.66	0.377	95.3	91.3	0.1	0.66	0.386
$(1, -1/2, -1/2)$															
10/10/10	90.4	28.2	98.1	0.57	0.672	95.5	88.2	0.7	0.59	0.380	95.5	87.5	0.9	0.59	0.391
30/30/30	93.6	62.1	8.5	0.35	0.634	95.2	92.9	0.0	0.35	0.454	95.1	92.9	0.0	0.35	0.461
30/10/10	92.7	47.7	46.8	0.44	0.581	95.4	92.5	0.1	0.44	0.383	95.4	86.7	0.1	0.44	0.391
30/20/10	93.0	52.0	31.3	0.41	0.591	95.3	86.4	0.0	0.41	0.402	95.3	86.4	0.0	0.41	0.410
$(-1, 1/2, 2)$															
10/10/10	89.1	28.2	99.2	1.06	0.736	95.4	86.3	1.7	1.09	0.424	95.4	86.3	2.0	1.09	0.436
30/30/30	93.2	62.3	21.6	0.64	0.680	95.1	91.9	0.0	0.65	0.493	95.1	91.9	0.0	0.65	0.501
30/10/10	87.8	42.4	97.2	0.98	0.787	95.3	86.3	4.5	1.02	0.481	95.2	86.3	6.3	1.02	0.492
30/20/10	87.3	35.6	96.8	0.96	0.800	95.3	88.9	8.2	1.01	0.493	95.2	84.7	10.5	1.00	0.505
$(-2, 1, 2)$															
10/10/10	91.0	28.2	98.6	1.41	0.624	95.6	86.7	0.2	1.44	0.337	95.5	86.4	0.2	1.44	0.347
30/30/30	93.8	62.3	6.1	0.85	0.590	95.2	92.7	0.0	0.85	0.416	95.1	92.7	0.0	0.85	0.423
30/10/10	91.3	47.7	92.6	1.19	0.628	95.4	86.3	0.3	1.21	0.381	95.4	63.3	0.4	1.21	0.390
30/20/10	91.0	51.8	90.7	1.14	0.636	95.4	86.3	0.5	1.17	0.399	95.3	86.3	0.6	1.17	0.408
$(1/3, 1/2, 3)$															
10/10/10	86.1	28.2	99.5	0.94	0.861	95.3	87.3	15.4	0.99	0.550	95.2	87.3	19.9	0.99	0.562
30/30/30	91.1	41.4	55.7	0.84	0.805	95.0	90.5	0.4	0.86	0.607	94.9	90.5	0.6	0.86	0.615
30/10/10	83.4	17.9	99.3	1.35	0.866	95.2	87.1	31.1	1.44	0.553	95.1	86.7	33.0	1.44	0.566
30/20/10	82.4	20.2	98.3	1.33	0.878	95.1	86.7	32.8	1.43	0.559	95.0	86.5	34.5	1.43	0.572

Table 5: Results of the evaluation measures for variants-3, 4 of the Wald CI, for $k = 3$. Confidence level $1 - \alpha = 95\%$. Distance d, defined by (25), is aimed at summarily differentiating between the performance of variant-3 and variant-4. The value of d is the same for all cases where the n_i's are equal for $k = 3$ binomial populations.

Method: **Wald**-95%		adjusted													
$(\beta_1, \beta_2, \beta_3)$		variant-3							variant-4						
$n_1/n_2/n_3$	d	R_{mean} (%)	R_{min} (%)	$R_{93\%}$ (%)	L_{mean}	MNR_{mean} (%)	DNR_{mean} (%)	Q_{mean}	R_{mean} (%)	R_{min} (%)	$R_{93\%}$ (%)	L_{mean}	MNR_{mean} (%)	DNR_{mean} (%)	Q_{mean}
$(1/3, 1/3, 1/3)$															
10/10/10	(0.00)	97.0	92.2	0.1	0.30	0.57	2.45	0.188	97.0	92.2	0.1	0.30	0.57	2.45	0.188
30/30/30		95.6	93.6	0.0	0.17	1.62	2.74	0.371	95.6	93.6	0.0	0.17	1.62	2.74	0.371
30/10/10	(0.23)	96.7	92.3	0.0	0.26	0.82	2.45	0.250	96.7	91.8	0.0	0.26	0.77	2.52	0.234
30/20/10	(0.27)	96.4	92.3	0.0	0.23	1.06	2.53	0.296	96.3	92.5	0.0	0.23	1.03	2.65	0.281
$(1, 1, -1)$															
10/10/10	(0.00)	97.0	92.0	0.0	0.91	0.56	2.47	0.186	97.0	92.0	0.0	0.91	0.56	2.47	0.186
30/30/30		95.6	94.0	0.0	0.51	1.62	2.75	0.370	95.6	94.0	0.0	0.51	1.62	2.75	0.370
30/10/10	(0.23)	96.7	92.3	0.0	0.79	0.80	2.49	0.244	96.7	93.4	0.0	0.79	0.75	2.58	0.226
30/20/10	(0.27)	96.4	93.1	0.0	0.69	1.05	2.55	0.292	96.3	92.7	0.0	0.69	1.01	2.68	0.274
$(1, -1/2, -1/2)$															
10/10/10	(0.41)	96.9	91.5	0.0	0.64	0.71	2.34	0.231	96.9	90.1	0.5	0.63	0.51	2.64	0.162
30/30/30		95.6	92.9	0.0	0.36	1.83	2.55	0.418	95.7	90.9	0.1	0.36	1.44	2.89	0.333
30/10/10	(0.08)	96.5	92.9	0.0	0.47	1.10	2.39	0.315	96.5	92.9	0.0	0.47	1.06	2.41	0.306
30/20/10	(0.21)	96.2	93.2	0.0	0.43	1.31	2.47	0.347	96.2	92.1	0.0	0.43	1.21	2.57	0.320
$(-1, 1/2, 2)$															
10/10/10	(0.53)	96.9	90.4	0.1	1.18	0.80	2.30	0.258	96.8	90.8	0.6	1.16	0.46	2.76	0.142
30/30/30		95.6	93.0	0.0	0.66	2.01	2.40	0.456	95.7	91.9	0.2	0.66	1.35	2.95	0.315
30/10/10	(0.66)	96.7	91.4	0.6	1.09	1.05	2.22	0.322	96.5	91.8	0.3	1.06	0.53	2.93	0.154
30/20/10	(0.69)	96.6	89.7	3.2	1.07	1.13	2.23	0.336	96.4	92.5	0.2	1.04	0.56	3.03	0.156
$(-2, 1, 2)$															
10/10/10	(0.27)	96.9	90.9	0.1	1.56	0.64	2.42	0.208	96.9	91.5	0.0	1.56	0.53	2.53	0.174
30/30/30		95.6	93.1	0.0	0.88	1.67	2.71	0.381	95.7	90.3	0.0	0.88	1.51	2.82	0.348
30/10/10	(0.37)	96.7	91.4	0.0	1.30	0.96	2.38	0.287	96.5	91.5	0.1	1.28	0.84	2.61	0.244
30/20/10	(0.44)	96.5	92.5	0.0	1.23	1.09	2.43	0.311	96.3	92.8	0.1	1.22	0.93	2.77	0.251
$(1/3, 1/2, 3)$															
10/10/10	(0.77)	96.8	89.8	14.0	1.55	1.16	2.07	0.359	96.7	92.3	0.1	1.04	0.33	2.90	0.101
30/30/30		95.5	92.9	0.0	0.87	2.53	1.92	0.568	96.0	94.7	0.0	0.88	1.18	2.85	0.293
30/10/10	(0.78)	96.7	88.9	14.9	1.54	1.20	2.09	0.364	96.7	94.4	0.0	1.49	0.33	2.95	0.102
30/20/10	(0.79)	96.6	89.0	15.9	1.52	1.25	2.12	0.372	96.7	94.4	0.0	1.47	0.34	3.00	0.102

Table 6: Results of some of the evaluation measures for the classic Newcombe-Zou, Peskun and score CI methods, for $k = 3$. Confidence level $1 - \alpha = 95\%$.

Method: **score**-95%	classic					classic					classic				
$(\beta_1, \beta_2, \beta_3)$	Newcombe-Zou					Peskun					score				
$n_1/n_2/n_3$	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}
	(%)	(%)	(%)			(%)	(%)	(%)			(%)	(%)	(%)		
$(1/3, 1/3, 1/3)$															
10/10/10	95.3	89.0	5.7	0.27	0.281	97.4	92.2	0.2	0.32	0.427	94.3	92.1	7.1	0.27	0.478
30/30/30	95.2	90.3	0.3	0.16	0.364	97.3	92.3	0.0	0.19	0.459	94.8	92.3	0.0	0.16	0.476
30/10/10	95.3	89.5	0.8	0.24	0.133	97.6	93.6	0.0	0.29	0.429	95.0	92.2	0.0	0.24	0.449
30/20/10	95.3	88.0	0.4	0.21	0.319	97.5	93.9	0.0	0.26	0.436	95.1	93.6	0.0	0.22	0.434
$(1, 1, -1)$															
10/10/10	95.2	86.9	5.7	0.81	0.279	97.4	91.7	0.1	0.94	0.427	94.4	91.7	6.7	0.82	0.479
30/30/30	95.2	91.4	0.4	0.49	0.363	97.3	93.1	0.0	0.57	0.459	94.8	91.9	0.0	0.49	0.476
30/10/10	95.3	89.6	0.8	0.72	0.293	97.6	94.0	0.0	0.87	0.426	94.9	91.9	0.0	0.73	0.446
30/20/10	95.3	91.0	0.4	0.64	0.314	97.5	94.2	0.0	0.77	0.432	95.1	93.7	0.0	0.65	0.431
$(1, -1/2, -1/2)$															
10/10/10	95.3	87.8	1.5	0.57	0.255	97.4	93.5	0.0	0.67	0.389	95.1	92.3	0.2	0.58	0.357
30/30/30	95.2	91.2	0.1	0.35	0.346	97.4	94.1	0.0	0.40	0.437	94.9	93.3	0.0	0.35	0.444
30/10/10	95.2	89.9	0.6	0.44	0.337	97.3	93.7	0.0	0.51	0.444	94.4	92.4	0.1	0.44	0.480
30/20/10	95.2	89.8	0.3	0.41	0.341	97.4	94.7	0.0	0.47	0.442	94.6	93.2	0.0	0.41	0.477
$(-1, 1/2, 2)$															
10/10/10	95.3	90.4	1.4	1.05	0.240	97.4	93.8	0.0	1.25	0.364	95.3	92.5	0.1	1.07	0.316
30/30/30	95.2	91.3	0.1	0.64	0.338	97.4	94.6	0.0	0.75	0.425	95.1	93.9	0.0	0.64	0.410
30/10/10	95.3	91.4	0.7	0.98	0.238	97.5	94.2	0.0	1.22	0.358	95.5	90.9	0.1	0.99	0.276
30/20/10	95.3	91.2	0.5	0.96	0.236	97.6	94.4	0.0	1.22	0.356	95.5	90.3	0.1	0.97	0.269
$(-2, 1, 2)$															
10/10/10	95.3	88.7	1.5	1.40	0.266	97.4	93.7	0.0	1.64	0.409	95.0	91.4	0.1	1.43	0.434
30/30/30	95.2	90.6	0.1	0.85	0.355	97.4	94.4	0.0	0.99	0.449	94.8	93.5	0.0	0.85	0.474
30/10/10	95.3	90.0	0.6	1.18	0.301	97.5	94.1	0.0	1.42	0.421	95.2	93.2	0.0	1.20	0.402
30/20/10	95.3	91.3	0.3	1.14	0.301	97.5	94.6	0.0	1.38	0.418	95.4	92.2	0.0	1.15	0.391
$(1/3, 1/2, 3)$															
10/10/10	95.3	89.7	1.6	1.36	0.211	97.4	93.9	0.0	1.67	0.303	95.5	86.6	0.3	1.37	0.223
30/30/30	95.2	91.9	0.2	0.84	0.321	97.3	94.6	0.0	1.01	0.392	95.3	92.6	0.0	0.84	0.337
30/10/10	95.3	90.0	1.6	1.35	0.210	97.6	94.3	0.0	1.75	0.309	95.5	86.6	0.5	1.36	0.221
30/20/10	95.4	90.0	1.1	1.34	0.209	97.6	94.5	0.0	1.76	0.308	95.5	87.5	0.3	1.35	0.217

6 CIs for Examples 1 and 2

We now analyse the two examples introduced in Section 2. Their (β_i,n_i)-configurations were not treated in our previous simulation study. Although the parameter settings in Example 1 can be considered similar to one of the cases reported in Table 8, we decided to carry out a simulation study using the particular settings of these two examples in order to compare the performance of the different CI methods and, therefore, decide which method is the best for each particular (β_i,n_i)-configuration.

In Example 1, there are $k = 4$ binomial populations and the same large number of samples from each population ($n_i = 30/30/30/30$). The results of our simulation were similar to those depicted in Table 8 for the second β_i-configuration and large sample sizes ($n_i = 20/20/20/20$). Thus, based on the four evaluation rules, we can conclude that the best methods for the particular scenario of Example 1 are variant-1 and variant-2 of the Wald method, since it is expected that they provide slightly conservative and less wide CIs with 0.0% liberal failures and satisfactory location. These two methods yield the intervals $]-0.3806, 0.2516[$ and $]-0.3808, 0.2516[$, respectively, as two-sided 95% CIs for $p_1 - p_2 - p_3 + p_4$ and thus conduct us to the no-rejection of H_0 in (3).

In Example 2, we have a (β_i,n_i)-configuration with $k = 7$ proportions and moderate and large samples sizes. Regarding specificity, the extremal proportions observed in four of the studies in the meta-analysis, suggest that the real specificity of the diagnostic test under evaluation is very high. Since the calculation of the exact coverage probabilities R is a computationally intensive process, we used the Monte Carlo method in this example. The simulation we performed was based on 1000 sets of 200000 k-samples from k binomial populations $B(n_i, p_i)$, $i = 1, \ldots, 7$. The simulation included only scenarios with very high proportions randomly generated from uniform distributions within [0.95,1]. Each replication provided an estimate of R, L_{mean}, MNR, DNR and Q. Estimates of each of the evaluation measures were obtained by calculating the mean of the 1000 corresponding values obtained during the replication process. Table 7 summarizes the averaged results. Although the score method produces slightly conservative and not too wide CIs, their locations tend to be completely non-satisfactory. Regarding the coverage probability R, the second best result is provided by variant-4 of the Wald method, which yielded CIs with 0.0% liberal failures and mesially satisfactory location. Remark that variant-4 is the best performer among the adjusted Wald CIs. According to the results we obtained, the new variant we propose seems to be the best method for the scenario of Example 2, yielding the interval $]0.888, 0.991[$ as a two-sided 95% CI for the pooled specificity of the seven studies considered. By way of curiosity, note that the score method yields the interval $]0.942, 0.988[$, which in fact seems to be too mesially located.

7 Conclusions

In this paper we have presented a systematic comparison of the performance of several approximate CI methods for a linear combination of binomial proportions of $k \geq 2$ independent populations, when the population proportions are estimated under the maximum likelihood saturated model (classic Wald CI)

Table 7: Evaluation of the eight CIs for Example 2.

CI-variant	R_{mean} (%)	R_{min} (%)	$R_{93\%}$ (%)	L_{mean}	MNR_{mean} (%)	DNR_{mean} (%)	Q_{mean}	Interval location
Wald-0	83.2	66.5	100	0.037	16.5	0.29	0.983	m.n.s. and d.n.s. - much too distal
Wald-1	98.8	98.0	0.0	0.050	0.02	1.21	0.014	m.n.s. and d.n.s. - much too mesial
Wald-2	98.8	97.9	0.0	0.050	0.03	1.15	0.023	m.n.s. and d.n.s. - much too mesial
Wald-3	98.9	98.1	0.0	0.091	0.00	1.15	0.000	m.n.s. and d.n.s. - much too mesial
Wald-4	97.6	96.6	0.0	0.094	0.00	2.45	0.000	only m.s.
Newcombe-Zou-0	90.6	89.8	99.9	0.064	0.00	9.41	0.000	m.n.s. and d.n.s. - much too mesial
Peskun-0	100.0	99.9	0.0	0.115	0.00	0.01	0.000	m.n.s. and d.n.s.
score-0	95.9	95.1	0.0	0.053	0.00	4.13	0.000	m.n.s. and d.n.s. - much too mesial

m.s. - mesially satisfactory m.n.s. - mesially non-satisfactory d.n.s. - distally non-satisfactory

and subjected to different constraints (classic Newcombe-Zou, classic Peskun and classic score CIs). Four adjusted variants based on the parametric family of shrinkage estimators $(X_i + h_i)/(n_i + 2h_i)$, $h_i > 0$, were considered for each of these classic CIs. Each choice for h_i establishes a different variant of the method for constructing approximate CIs for $L = \sum_{i=1}^{k} \beta_i p_i$. The parameters h_i suggested in the literature for the herein designated variants-$1, 2, 3$ do not take into account the effect of the estimate found for p_i on the linear combination L. To overcome this lack, we have proposed a new choice for h_i based on the h_i used in variant-3, which establishes a new variant of each of the above mentioned CI methods, herein designated variant-4, that has the advantage of also handling extremal observations. This new variant is defined in terms of the weights β_i of the proportions p_i and is balanced by the sample sizes n_i of the corresponding populations in the linear combination. The performance of the CI methods was evaluated via simulation by the calculation of various statistical measures based on the exact coverage probability, expected interval length and mesial and distal non-coverage probabilities (R_{mean}, R_{min}, $R_{93\%}$, L_{mean}, MNR_{mean} and DNR_{mean}). The interval location was characterized through the Q_{mean} ratio and also the MNR_{mean} and DNR_{mean} values, individually. The two latter measures are of particular interest when an imbalance between the mesial and distal non-coverage probabilities can potentially occur, namely, for settings involving small and large samples from different populations or for cases involving extremal observations (an especially appropriate situation for variants-$3, 4$). The effect of the continuity correction proposed by Haber on the CIs was also studied.

We found that the CIs obtained through the classic Wald and the adjusted score (variants-$1, 2, 3, 4$) methods had the poorest performance in all the parameter settings (n_i, β_i) considered. Variants-$1, 2$ of the adjusted Wald method showed good performance in terms of coverage probability and were the most consistent in producing satisfactorily located CIs, in the sense that an equilibrium between the MNR_{mean} and DNR_{mean} values exists. In spite of the fact that variant-3 and variant-4 have yielded almost identical R_{mean} values and are very conservative, variant-4 exhibited the best performance in terms of the $R_{93\%}$, R_{min} and L_{mean}. For these two variants, the location of the CIs was mesially satisfactory for almost all settings, having a very low probability of being distally located (lower for variant-4). The classic Newcombe-Zou CI method was the

best performer in terms of R_{mean} and L_{mean} and its $R_{93\%}$ performance was also good for some settings. Its Q_{mean} values showed a clear tendency of this method to produce too mesially located CIs. According to the values of MNR_{mean} and DNR_{mean}, these intervals are mesially and distally non-satisfactory. The classic variant of the Peskun CI yielded the lowest percentages of liberal failures but was too conservative and its interval locations were always mesially and distally non-satisfactory. The classic score CI method also had good performance in terms of $R_{93\%}$ (except for small samples) and yielded one of the highest accuracy CIs. In terms of interval location, it switched from satisfactory (mesially and distally satisfactory) to too mesial (mesially and distally non-satisfactory) when an imbalance existed between the MNR_{mean} and DNR_{mean} values. Comparing all variants for the settings considered, the confidence intervals of the Wald method obtained from the new variant exhibit similar performance in terms of the exact coverage probability and the expected interval length and show improved performance when imbalances among the k populations are observed between the weight of each proportion in the linear combination and the sample size drawn from the population. They have satisfactory mesial locations and a very low probability of being distally located, which is consistent with the capability of this variant to handle extremal observations. The effect of applying a continuity correction to the various methods for $k = 3$ and $k = 4$ is a slight reduction of the liberal failures, which is particularly small for $k = 4$.

In view of our conclusions resulting from the analysis of all parameter settings and the two examples herein considered, it is obvious that different scenarios can lead to different choices among competing CI procedures. There is no procedure that outperforms all others in all (n_i, β_i)-configurations.

In general, the classic Newcombe-Zou CI will be the best performer for $k = 3$ and $k = 4$ binomial populations, if the selection of the best method is based on the R_{mean} and L_{mean} (selection criterion considered by some researchers). Nevertheless, its performance in terms of $R_{93\%}$ is not that good and its location is mesially and distally non-satisfactory. We address the selection of the best method based on four rules, preferring methods with the smallest number of liberal failures as suggested by some other researchers. The new variant we propose for the adjusted Wald method, variant-4, exhibits good behaviour and stability across most of the distinct settings considered and, due to its performance in terms of $R_{93\%}$, would be the recommended method for $k = 3$ and $k = 4$. This variant seems to be particularly adequate to contexts where mesially shifted estimates are preferred.

References

A An approximation of the centre of the classic score CI

Let $Z_S^2/(\hat{L}-\lambda_0)=C$ in equation (8). The next step will be solving (8) in order to C, which is impossible to do analytically. Although numerical procedures for obtaining the values of C seems to be computationally expeditious, they have the inconvenience of not providing explicit expressions either for the bounds of confidence limits in (9) or for the centre of the CI. In order to establish approximate formulas for them, representations of $R_i=\sqrt{(\beta_i C+n_i)^2-4\beta_i\, n_i\hat{p}_i C}$ using a finite number of initial terms in its MacLaurin series expansion could be considered. Indeed, calculating the terms of the series of R_i up to the order $O(1/n_i^2)$ around $C=0$, we obtain, after some algebraic simplifications,

$$R_i \approx \begin{cases} n_i+\beta_i b_i C+\dfrac{2\beta_i^2\hat{p}_i(1-\hat{p}_i)}{n_i}C^2-\dfrac{2\beta_i^3\hat{p}_i(1-\hat{p}_i)b_i}{n_i^2}C^3 & \text{, if } \ 0<\hat{p}_i<1 \\ -(n_i+b_i\beta_i C) & \text{, if } (\hat{p}_i=0\vee\hat{p}_i=1)\wedge(b_i\beta_i C<0) \end{cases} \tag{26}$$

with $b_i=1-2\hat{p}_i$ (**?**). Note that the second branch in (26) addresses the extremal cases $\hat{p}_i=0,1$. Concretely, since $C(\hat{L}-\lambda_0)=Z_S^2>0$, the boundary condition for those extremal cases can be expressed in terms of the set $\mathcal{A}_i$ defined in (17). In these circumstances, if we set $Z_S^2=z_{\alpha/2}^2$, then $C=z_{\alpha/2}^2/(\hat{L}-\lambda_0)$. If we now substitute C into (26), we will obtain an approximate second degree equation in terms of λ_0 from (8), after some algebraic simplifications. The solutions of this equation determine the confidence bounds (9) of the classic score CI, whose centre can be expressed by

$$\hat{L}+\frac{B^*\, z_{\alpha/2}^2}{2\left(N^*+z_{\alpha/2}^2\right)}+\frac{z_{\alpha/2}^2}{2}\,\frac{\displaystyle\sum_{i=1}^{k}\frac{\beta_i^3\hat{p}_i(1-\hat{p}_i)b_i}{n_i^2}}{\displaystyle\sum_{i=1}^{k}\frac{\beta_i^2\,\hat{p}_i\,(1-\hat{p}_i)}{n_i}}, \tag{27}$$

Remark that the fulfilment of the conditions $\lambda_0<\hat{L}$ and $\lambda_0>\hat{L}$ implies that the boundary condition in $\mathcal{A}_i$ will affect the value of the lower and upper bounds of the CI in a different manner. Indeed, for the lower bound of the CI (9), the condition $\lambda_0<\hat{L}$ in $\mathcal{A}_i$ holds and therefore $\mathcal{A}_i$ will be given by $\{x_i:\ (x_i=0\wedge\beta_i<0)\vee(x_i=n_i\wedge\beta_i>0)\}$. A similar reasoning holds for the upper bound, where $\mathcal{A}_i$ will be given by $\{x_i:\ (x_i=0\wedge\beta_i>0)\vee(x_i=n_i\wedge\beta_i<0)\}$. Detailed expressions of the CIs when R_i in (26) is approximated by terms up to the orders $O(1/n_i)$ and $O(1/n_i^2)$ can be found in **?**, expressions (9) and (10).

B Simulation results for $k=4$

Table 8: Results of some of the evaluation measures for variants-0, 1, 2 of the Wald CI, for $k = 4$. Confidence level $1 - \alpha = 95\%$.

Method: **Wald**-95%	classic					adjusted									
$(\beta_1, \beta_2, \beta_3, \beta_4)$	variant-0					variant-1					variant-2				
$n_1/n_2/n_3/n_4$	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}
	(%)	(%)	(%)			(%)	(%)	(%)			(%)	(%)	(%)		
$(1/4, 1/4, 1/4/1/4)$															
10/10/10/10	92.4	67.6	88.9	0.24	0.577	95.3	92.5	0.0	0.24	0.362	95.2	92.4	0.0	0.24	0.371
20/20/20/20	93.8	84.3	3.1	0.17	0.563	95.1	93.9	0.0	0.17	0.408	95.1	93.8	0.0	0.17	0.414
20/20/10/10	92.7	72.7	72.8	0.21	0.578	95.2	92.7	0.0	0.21	0.390	95.1	92.7	0.0	0.21	0.397
20/15/10/5	90.3	63.1	96.2	0.23	0.587	95.3	91.2	0.5	0.24	0.368	95.2	91.2	0.7	0.24	0.368
$(-1, 1, -1, 1)$															
10/10/10/10	92.4	66.6	88.9	0.95	0.577	95.3	93.2	0.0	0.96	0.362	95.2	93.2	0.0	0.96	0.371
20/20/20/20	93.8	85.0	3.2	0.69	0.563	95.1	93.6	0.0	0.69	0.408	95.1	93.5	0.0	0.69	0.414
20/20/10/10	92.7	79.2	72.4	0.83	0.576	95.2	92.6	0.0	0.84	0.389	95.1	92.6	0.0	0.84	0.396
20/15/10/5	90.3	67.9	95.5	0.94	0.584	95.3	90.7	0.5	0.97	0.368	95.2	90.6	0.7	0.97	0.377
$(1/3, 1/3, 1/3, 1)$															
10/10/10/10	89.9	56.4	98.8	0.54	0.730	94.9	89.0	4.3	0.55	0.495	94.8	89.0	6.1	0.55	0.504
20/20/20/20	92.6	77.3	61.9	0.39	0.696	94.9	92.1	0.1	0.40	0.528	94.8	92.1	0.1	0.40	0.535
20/20/10/10	89.0	62.4	97.6	0.51	0.760	94.8	88.8	14.1	0.53	0.528	94.7	88.8	20.3	0.53	0.537
20/15/10/5	80.1	42.1	98.5	0.62	0.807	94.8	82.7	39.5	0.69	0.460	94.6	82.7	41.1	0.69	0.472
$(-3, -1, 1, 3)$															
10/10/10/10	91.1	53.5	98.4	2.09	0.614	95.1	89.6	0.1	2.14	0.406	95.0	89.6	0.1	2.14	0.414
20/20/20/20	93.2	75.0	22.7	1.53	0.601	95.0	89.0	0.0	1.54	0.444	94.9	88.9	0.0	1.54	0.451
20/20/10/10	91.5	66.3	91.5	1.83	0.623	95.0	89.3	0.7	1.86	0.433	94.9	89.3	0.8	1.86	0.441
20/15/10/5	85.6	43.6	95.3	2.15	0.646	95.1	84.1	34.3	2.29	0.410	94.9	83.0	36.2	2.29	0.419
$(1/6, 1/3, 1/2, 3)$															
10/10/10/10	84.2	35.4	99.9	1.36	0.860	94.3	84.2	42.6	1.44	0.626	94.2	82.7	44.2	1.44	0.635
20/20/20/20	89.4	43.8	96.9	1.01	0.830	94.4	87.6	11.5	1.04	0.656	94.3	87.4	13.2	1.04	0.663
20/20/10/10	83.6	26.6	99.7	1.35	0.865	94.3	83.6	42.8	1.43	0.629	94.2	82.8	44.2	1.43	0.638
20/15/10/5	71.8	15.9	98.1	1.64	0.887	94.5	82.0	31.0	1.93	0.545	94.3	81.6	32.5	1.93	0.558
$(-1/2, 1/2, 1, 4)$															
10/10/10/10	86.2	38.5	99.8	1.88	0.831	94.5	85.9	42.6	1.97	0.599	94.4	85.8	45.8	1.97	0.609
20/20/20/20	90.6	61.6	95.0	1.40	0.797	94.6	89.4	2.6	1.42	0.628	94.5	89.4	3.3	1.42	0.635
20/20/10/10	85.7	38.5	99.5	1.86	0.838	94.5	85.6	43.1	1.96	0.606	94.3	85.0	45.8	1.96	0.615
20/15/10/5	74.6	21.7	99.3	2.28	0.873	94.6	82.0	33.4	2.62	0.520	94.4	82.0	34.7	2.62	0.534

Table 9: Results of the evaluation measures for variants-3, 4 of the Wald CI, for $k = 4$. Confidence level $1 - \alpha = 95\%$. Distance d, defined by (25), is aimed at summarily differentiating between the performance of variant-3 and variant-4. The value of d is the same for all cases where the n_i's are equal for $k = 4$ binomial populations.

Method: **Wald**-95%		adjusted													
$(\beta_1, \beta_2, \beta_3, \beta_4)$		variant-3							variant-4						
$n_1/n_2/n_3/n_4$	d	R_{mean} (%)	R_{min} (%)	$R_{93\%}$ (%)	L_{mean}	MNR_{mean} (%)	DNR_{mean} (%)	Q_{mean}	R_{mean} (%)	R_{min} (%)	$R_{93\%}$ (%)	L_{mean}	MNR_{mean} (%)	DNR_{mean} (%)	Q_{mean}
$(1/4, 1/4, 1/4, 1/4)$															
10/10/10/10	(0.00)	97.1	93.3	0.0	0.27	0.61	2.25	0.214	97.1	93.3	0.0	0.27	0.61	2.25	0.214
20/20/20/20		96.1	94.0	0.0	0.18	1.35	2.60	0.341	96.1	94.0	0.0	0.18	1.35	2.60	0.341
20/20/10/10	(0.17)	96.7	94.1	0.0	0.23	0.91	2.37	0.277	96.7	93.4	0.0	0.23	0.88	2.41	0.269
20/15/10/5	(0.28)	97.5	94.9	0.0	0.27	0.53	1.97	0.211	97.2	93.3	0.0	0.26	0.57	2.19	0.205
$(-1, 1, -1, 1)$															
10/10/10/10	(0.00)	97.1	93.8	0.0	1.06	0.61	2.25	0.213	97.1	93.8	0.0	1.06	0.61	2.25	0.213
20/20/20/20		96.1	94.1	0.0	0.73	1.34	2.60	0.340	96.1	94.1	0.0	0.73	1.34	2.60	0.340
20/20/10/10	(0.17)	96.7	94.1	0.0	0.91	0.90	2.37	0.276	96.7	93.1	0.0	0.90	0.88	2.40	0.268
20/15/10/5	(0.28)	97.5	93.4	0.0	1.08	0.53	1.97	0.211	97.3	93.6	0.0	1.06	0.56	2.18	0.206
$(1/3, 1/3, 1/3, 1)$															
10/10/10/10	(0.58)	97.0	92.7	0.0	0.60	0.91	2.16	0.297	96.9	91.8	0.3	0.59	0.46	2.63	0.148
20/20/20/20		95.9	93.7	0.0	0.41	1.86	2.27	0.450	96.1	92.5	0.1	0.41	1.06	2.88	0.269
20/20/10/10	(0.66)	96.7	92.6	0.4	0.57	1.12	2.20	0.337	96.6	91.8	0.5	0.56	0.52	2.86	0.154
20/15/10/5	(0.74)	97.8	93.0	0.0	0.77	0.28	1.89	0.131	97.0	91.5	1.3	0.70	0.17	2.84	0.057
$(-3, -1, 1, 3)$															
10/10/10/10	(0.40)	97.0	93.2	0.0	2.35	0.76	2.27	0.251	97.0	92.3	0.0	2.33	0.50	2.45	0.170
20/20/20/20		95.9	93.9	0.0	1.61	1.52	2.55	0.374	96.1	93.0	0.0	1.61	1.15	2.77	0.293
20/20/10/10	(0.45)	96.6	93.4	0.0	2.01	1.07	2.34	0.315	96.6	93.0	0.0	1.99	0.77	2.63	0.225
20/15/10/5	(0.58)	97.7	93.4	0.0	2.55	0.52	1.81	0.222	97.0	92.0	1.2	2.40	0.46	2.58	0.150
$(1/6, 1/3, 1/2, 3)$															
10/10/10/10	(0.82)	96.6	89.3	18.4	1.56	1.39	2.02	0.407	96.8	95.1	0.0	1.51	0.3	2.9	0.101
20/20/20/20		95.6	91.6	1.2	1.08	2.49	1.86	0.572	96.2	95.1	0.0	1.08	0.9	2.9	0.236
20/20/10/10	(0.83)	96.5	89.1	19.2	1.55	1.43	2.05	0.412	96.7	94.8	0.0	1.49	0.3	2.9	0.102
20/15/10/5	(0.84)	97.8	91.9	0.3	2.16	0.06	2.18	0.027	97.1	92.5	0.0	1.93	0.0	2.9	0.005
$(-1/2, 1/2, 1, 4)$															
10/10/10/10	(0.77)	96.7	91.1	14.5	2.14	1.2	2.1	0.377	96.8	92.3	0.1	2.08	0.4	2.8	0.111
20/20/20/20		95.7	92.6	0.0	1.48	2.3	1.9	0.546	96.2	91.9	0.1	1.48	0.9	2.9	0.240
20/20/10/10	(0.78)	96.6	90.1	18.3	2.12	1.3	2.1	0.386	96.7	93.2	0.0	2.05	0.4	2.9	0.112
20/15/10/5	(0.82)	97.8	92.4	0.1	2.93	0.1	2.1	0.055	97.1	93.4	0.0	2.63	0.0	2.9	0.013

Table 10: Results of some of the evaluation measures for the classic Newcombe-Zou, Peskun and score CI methods, for $k = 4$. Confidence level $1 - \alpha = 95\%$.

Method: **score**-95%	classic					classic					classic				
$(\beta_1, \beta_2, \beta_3, \beta_4)$	Newcombe-Zou					Peskun					score				
$n_1/n_2/n_3/n_4$	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}
	(%)	(%)	(%)			(%)	(%)	(%)			(%)	(%)	(%)		
$(1/4, 1/4, 1/4, 1/4)$															
10/10/10/10	95.2	89.6	4.8	0.24	0.304	97.7	92.9	0.0	0.28	0.443	93.8	91.7	6.5	0.24	0.488
20/20/20/20	95.2	91.5	0.6	0.17	0.351	97.6	93.6	0.0	0.20	0.462	94.5	92.0	0.1	0.17	0.484
20/20/10/10	95.2	91.6	0.5	0.21	0.322	97.7	94.2	0.0	0.25	0.450	94.4	93.2	0.0	0.21	0.480
20/15/10/5	95.2	90.9	1.2	0.23	0.297	97.8	94.9	0.0	0.29	0.429	95.1	93.2	0.0	0.24	0.423
$(-1, 1, -1, 1)$															
10/10/10/10	95.2	88.6	4.8	0.94	0.303	97.7	92.9	0.0	1.13	0.443	93.8	91.8	6.4	0.94	0.487
20/20/20/20	95.2	91.3	0.5	0.69	0.351	97.6	93.9	0.0	0.82	0.463	94.5	92.1	0.1	0.69	0.485
20/20/10/10	95.2	91.3	0.7	0.82	0.323	97.7	94.2	0.0	1.00	0.450	94.4	93.1	0.0	0.83	0.481
20/15/10/5	95.2	90.3	1.2	0.94	0.299	97.8	94.9	0.0	1.17	0.430	95.1	92.1	0.0	0.96	0.425
$(1/3, 1/3, 1/3, 1)$															
10/10/10/10	95.2	90.3	1.4	0.53	0.246	97.6	94.2	0.0	0.65	0.372	95.3	93.4	0.0	0.54	0.312
20/20/20/20	95.2	91.1	0.2	0.39	0.310	97.6	94.4	0.0	0.47	0.410	95.2	39.4	0.0	0.94	0.381
20/20/10/10	95.3	90.5	0.6	0.51	0.242	97.7	94.7	0.0	0.64	0.365	95.4	94.1	0.0	0.52	0.292
20/15/10/5	95.3	91.7	2.0	0.63	0.164	97.8	94.2	0.0	0.87	0.291	95.6	92.5	0.0	0.65	0.191
$(-3, -1, 1, 3)$															
10/10/10/10	95.3	90.4	1.0	2.09	0.264	97.7	94.4	0.0	2.52	0.409	95.0	92.4	0.0	2.12	0.431
20/20/20/20	95.2	90.3	0.1	1.53	0.324	97.7	94.5	0.0	1.83	0.438	94.7	93.5	0.0	1.53	0.476
20/20/10/10	95.3	90.7	0.4	1.82	0.289	97.7	94.7	0.0	2.23	0.417	95.2	93.7	0.0	1.85	0.410
20/15/10/5	95.3	91.1	1.1	2.17	0.246	97.8	94.9	0.0	2.84	0.375	95.6	92.5	0.0	2.23	0.325
$(1/6, 1/3, 1/2, 3)$															
10/10/10/10	95.3	91.5	1.5	1.37	0.210	97.5	94.8	0.0	1.73	0.307	95.5	91.9	0.1	1.37	0.224
20/20/20/20	95.2	91.7	0.3	1.01	0.285	97.5	94.1	0.0	1.25	0.369	95.4	93.7	0.0	1.02	0.301
20/20/10/10	95.3	92.1	1.5	1.36	0.210	97.6	94.7	0.0	1.77	0.310	95.5	91.7	0.2	1.37	0.222
20/15/10/5	95.5	88.7	3.9	1.74	0.117	97.8	93.5	0.0	2.50	0.218	95.6	88.7	1.2	1.75	0.122
$(-1/2, 1/2, 1, 4)$															
10/10/10/10	95.3	90.6	1.4	1.88	0.216	97.5	94.8	0.0	2.36	0.322	95.5	92.5	0.0	1.90	0.241
20/20/20/20	95.2	90.9	0.3	1.39	0.289	97.5	94.8	0.0	1.71	0.378	95.4	94.3	0.0	1.40	0.318
20/20/10/10	95.3	91.5	1.5	1.87	0.216	97.6	94.8	0.0	2.41	0.325	95.5	92.4	0.0	1.88	0.238
20/15/10/5	95.4	91.4	3.1	2.37	0.123	97.8	93.5	0.0	3.37	0.230	95.6	90.6	0.4	2.39	0.133

Table 11: Results of some of the evaluation measures for the classic Newcombe-Zou and classic score CI methods with continuity correction, for $k = 4$. Confidence level $1 - \alpha = 95\%$.

Method: **score**-95%	classic									
$(\beta_1, \beta_2, \beta_3, \beta_4)$	Newcombe-Zou (cc)					score (cc)				
$n_1/n_2/n_3/n_4$	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}	R_{mean}	R_{min}	$R_{93\%}$	L_{mean}	Q_{mean}
	(%)	(%)	(%)			(%)	(%)	(%)		
$(1/4, 1/4, 1/4, 1/4)$										
10/10/10/10	95.2	89.6	4.8	0.24	0.304	93.8	91.7	6.4	0.24	0.488
20/20/20/20	95.2	91.5	0.6	0.17	0.351	94.5	92.0	0.1	0.17	0.484
20/20/10/10	95.2	91.6	0.5	0.21	0.322	94.4	93.2	0.0	0.21	0.480
20/15/10/5	95.2	90.9	1.2	0.23	0.297	95.1	93.4	0.0	0.24	0.423
$(-1, 1, -1, 1)$										
10/10/10/10	95.2	88.6	4.8	0.94	0.303	93.8	91.8	6.3	0.95	0.487
20/20/20/20	95.2	91.3	0.5	0.69	0.351	94.5	92.1	0.1	0.69	0.485
20/20/10/10	95.2	91.3	0.7	0.82	0.323	94.4	93.1	0.0	0.83	0.481
20/15/10/5	95.2	90.3	1.2	0.94	0.299	95.1	92.1	0.0	0.96	0.425
$(1/3, 1/3, 1/3, 1)$										
10/10/10/10	95.2	90.3	1.3	0.53	0.246	95.3	93.4	0.0	0.54	0.312
20/20/20/20	95.2	91.1	0.2	0.39	0.310	95.4	94.1	0.0	0.52	0.292
20/20/10/10	95.3	90.6	0.6	0.51	0.242	95.4	94.1	0.0	0.52	0.292
20/15/10/5	95.3	91.7	1.9	0.63	0.164	95.6	92.5	0.0	0.65	0.191
$(-3, -1, 1, 3)$										
10/10/10/10	95.3	90.4	1.0	2.09	0.264	95.0	92.4	0.0	2.12	0.431
20/20/20/20	95.2	90.3	0.1	1.53	0.324	94.7	93.5	0.0	1.53	0.476
20/20/10/10	95.3	91.1	0.4	1.82	0.289	95.2	93.7	0.0	1.85	0.410
20/15/10/5	95.3	91.1	1.1	2.17	0.246	95.6	92.5	0.0	2.23	0.325
$(1/6, 1/3, 1/2, 3)$										
10/10/10/10	95.3	91.5	1.5	1.37	0.210	95.5	91.9	0.1	1.37	0.224
20/20/20/20	95.2	91.7	0.3	1.01	0.285	95.4	93.7	0.0	1.02	0.301
20/20/10/10	95.3	92.1	1.4	1.36	0.210	95.5	91.7	0.2	1.37	0.222
20/15/10/5	95.5	88.7	3.8	1.74	0.117	95.6	88.7	1.2	1.75	0.121
$(-1/2, 1/2, 1, 4)$										
10/10/10/10	95.3	90.6	1.4	1.88	0.216	95.5	92.5	0.0	1.90	0.241
20/20/20/20	95.2	90.9	0.3	1.39	0.289	95.4	94.3	0.0	1.40	0.317
20/20/10/10	95.3	91.5	1.5	1.87	0.216	95.5	92.4	0.0	1.88	0.238
20/15/10/5	95.4	91.4	3.0	2.37	0.123	95.6	90.6	0.4	2.39	0.133