Journal of Data Science 23 (4), 638–647
October 2025
DOI: 10.6339/25-JDS1168
Statistical Data Science

Analysis of Bilateral and Unilateral Data: A Comparative Review of Model-Based and MLE-Based Methods for the Homogeneity Test of Proportions

Xueqing Zhang and Chang-Xing Ma[1,*]
[1]*The State University of New York at Buffalo, Buffalo, NY, US*

Abstract

In many medical comparative studies, subjects may provide either bilateral or unilateral data. While numerous testing procedures have been proposed for bilateral data that account for the intra-class correlation between paired organs of the same individual, few studies have thoroughly explored combined correlated bilateral and unilateral data. Ma and Wang (2021) introduced three test procedures based on the maximum likelihood estimation (MLE) algorithm for general g groups. In this article, we employ a model-based approach that treats the measurements from both eyes of each subject as repeated observations. We then compare this approach with Ma and Wang's Score test procedure. Monte Carlo simulations demonstrate that the MLE-based Score test offers certain advantages under specific conditions. However, this model-based method lacks an explicit form for the test statistic, limiting its potential for further development of an exact test.

Keywords *correlated bilateral and unilateral data; generalized estimating equations; MLE-based test procedures; Rosner's model*

1 Introduction

In many medical comparative studies, subjects may produce data from paired organs, either bilateral (e.g., responses from a pair of ears, eyes, and hands) or unilateral (response from only one ear, eye, or hand). Specifically, in ophthalmologic studies, for bilateral cases, it is meaningful to assume that the information between the two eyes from the same subject is generally correlated. Rosner (1982) points out that the fundamental unit for statistical analysis in ophthalmologic studies is often the eye rather than the person. If an individual contributes two eyes worth of information to analysis, such as comparing intraocular pressures in persons in different age groups, their values are generally correlated. If the values are correlated, then methods of analysis in which each eye is considered an independent random variable are not valid. For the bilateral data, Rosner (1982) proposed an equal $\boldsymbol{R}$ model for testing whether the proportions of affected eyes are the same among the g groups of patients while accounting for the intra-person dependence. Dallal (1988), criticizing the appropriateness of Rosner's model for the case of a single binomially distributed eye-specific outcome variable, proposed an alternative approach based on compound multinomial sampling. Donner (1989) proposed an alternative approach based on a simple adjustment of the standard Pearson Chi-square test for homogeneity of proportions. Tang et al. (2008) investigated eight procedures for testing the equality of proportions between two

*Corresponding author. Email: cxma@buffalo.edu.

Received September 29, 2024; Accepted January 8, 2025

groups in correlated data. Empirical results show that tests based on the approximate unconditional method usually produce empirical type I error rates closer to the pre-chosen nominal level than their asymptotic tests. The maximum likelihood estimates under Rosner's model and three different methods (Likelihood Ratio test, Wald-type test, and Score test) are derived and investigated by Ma et al. (2015). In practice, Lu et al. (2022) employed these methodologies to compare the efficacy and safety of corneal refractive therapy lenses and vision-shaping treatment lenses for controlling myopia, providing evidence for the utility of statistical techniques in handling correlated outcomes in clinical trials. In an otolaryngology study, Mandel et al. (1982) demonstrated the importance of properly accounting for paired organ correlation when comparing treatments for acute otitis media. It should be noted that all the procedures mentioned above are applicable to bilateral data. Obviously, procedures that fail to utilize both unilateral and bilateral data would be less powerful. Pei et al. (2008) studied ten test statistics that utilize both the unilateral and bilateral data to test the equality of two proportions and found that both Rosner's and Wald-type statistics based on the dependence model and constrained maximum likelihood estimates perform satisfactorily for small to large samples. For general $g \geqslant 2$ groups, Ma and Wang (2021) extend Ma et al. (2015)'s work for bilateral data to combined bilateral and unilateral data. The result shows that the Score test has satisfactory type I error rates and powers.

This article addresses a gap in the literature by directly comparing Ma and Wang's Score test procedure with a model-based method that treats measurements from both eyes of each subject as repeated observations. Specifically, we explore testing the equality of general g proportions for combined bilateral and unilateral data under Rosner's model, while accounting for the correlation between eyes. To our knowledge, no previous study has directly compared these two methods in this context, making this work a valuable contribution. By doing so, we provide researchers with an opportunity to apply a Maximum Likelihood Estimation (MLE)-based approach, which requires less computational cost, and is well-suited for further development, making it an accessible and efficient option for analyzing complex data structures. The comparison not only enhances the understanding of both methods' strengths but also opens the door for further advancements and applications of the MLE-based method in similar settings.

We consider the observed data as in Table 1. Let π_i denote the probability of having a response in ith group, The equality of π_i among groups is of interest. Ma's Score test is briefly introduced in Section 2. In Section 3, we introduce Generalized estimating equations (GEE) as the theoretical underpinning of how the proposed method handles correlated binary outcomes and how to realize them in SAS, and evaluate its performance by comparing it to Ma's Score test. In Section 4, simulation studies are conducted to compare the performance of both methods. Comparisons are evaluated concerning type I error rates and powers through various configurations. Finally, we describe some findings from the result in Section 4 and give some conclusions.

2 Method

2.1 Ma and Wang's Score Test

Consider comparing g groups of individuals with m_i individuals that contribute two eyes and n_i individuals in the ith group that contribute one eye for the study. $M = \sum m_i$, $N = \sum n_i$, $i = 1, \ldots, g$. Let m_{ti} $(t = 0, 1, 2)$ be the number of subjects with t responses in the ith group who contribute two eyes, n_{ti} $(t = 0, 1)$ be the number of subjects with t responses in the ith group who contribute one eye. let M_t $(t = 0, 1, 2)$ and N_t $(t = 0, 1)$ be the number of subjects who have exactly t response, then $M_t = \sum_{i=1}^{g} m_{ti}$, and $N_t = \sum_{i=1}^{g} n_{ti}$.

Table 1: Frequencies of the number of affected eyes for persons in g groups.

Number of affected eyes	Group 1	2	...	g	Total
0	m_{01}	m_{02}	...	m_{0g}	M_0
1	m_{11}	m_{12}	...	m_{1g}	M_1
2	m_{21}	m_{22}	...	m_{2g}	M_2
Total	m_1	m_2	...	m_g	M
0	n_{01}	n_{02}	...	n_{0g}	N_0
1	n_{11}	n_{12}	...	n_{1g}	N_1
Total	n_1	n_2	...	n_g	N

Assuming equal dependence between two eyes of the same person across groups is proposed by Rosner (1982).

$$Pr(Z_{ijk} = 1) = \pi_i, \quad Pr(Z_{ijk} = 1 | Z_{ij,3-k} = 1) = R\pi_i,$$

where $Z_{ijk} = 1$ if the kth eye of jth individual in the ith group has a response at the end of the study and 0 otherwise. R is a positive constant that measures the dependence between two eyes of the same person. The correlation between two eyes of the same individual for the ith group can be calculated as

$$\rho_i = \frac{\pi_i}{1 - \pi_i}(R - 1).$$

The observed data can be written as $\tilde{D} = (m_{01}, m_{11}, m_{21}, \ldots, m_{0g}, m_{1g}, m_{2g}, n_{01}, n_{11}, \ldots, n_{0g}, n_{1g})$.

$$\begin{aligned}(m_{0i}, m_{1i}, m_{2i}) &\sim \text{Multinomial}\big(m_i, \big(R\pi_i^2 - 2\pi_i + 1, 2\pi_i(1 - R\pi_i), R\pi_i^2\big)\big),\\ n_{1i} &\sim \text{Binomial}(n_i, \pi_i).\end{aligned}$$

Then, the log-likelihood function is

$$\begin{aligned}l(\pi_1, \ldots, \pi_g; R) = &\sum_{i=1}^{g}\big[m_{0i}\log\big(R\pi_i^2 - 2\pi_i + 1\big) + m_{1i}\log\big(2\pi_i(1 - R\pi_i)\big) + m_{2i}\log\big(R\pi_i^2\big)\big]\\ &+ \sum_{i=1}^{g}\big[n_{0i}\log(1 - \pi_i) + n_{1i}\log(\pi_i)\big] + C. \qquad (1)\end{aligned}$$

Our goal is to test

$$H_0 : \pi_1 = \cdots = \pi_g = \pi \quad \text{v.s.} \quad H_1 : \text{some of the } \pi_i \text{ are unequal.}$$

Ma and Wang developed the constrained and unconstrained maximum likelihood estimates of π_i and R, and hence derived the Score test statistic T_{SC}. Let $U = (U_1, \ldots, U_g, 0) = (\frac{\partial l}{\partial \pi_1}, \ldots, \frac{\partial l}{\partial \pi_g}, 0)$, then the Score test statistic T_{SC} is

$$T_{SC} = U\,I(\pi, R)^{-1}U^T|_{\pi_1=\cdots=\pi_g=\hat{\pi}_{H_0}, R=\hat{R}_{H_0}}. \qquad (2)$$

After lengthy algebra calculation, T_{SC} can be simplified as

$$T_{SC} = \sum_{i=1}^{g} \frac{U^2}{I_{ii}} + \left(\sum_{i=1}^{g} \frac{I_{i,g+1} U_i}{I_{ii}} \right)^2 \left(I_{g+1,g+1} - \sum_{j=1}^{g} \frac{I_{j,g+1}^2}{I_{jj}} \right)^{-1}, \tag{3}$$

where

$$\begin{aligned}
I_{ii} &= E\left(-\frac{\partial^2 l}{\partial \pi_i^2} \right) = \frac{2m_i(2R^2\pi_i^2 - R\pi_i^2 - 2R\pi_i + 1)}{\pi_i(R\pi_i^2 - 2\pi_i + 1)(1 - R\pi_i)} + \frac{n_i}{\pi_i(1 - \pi_i)}, \\
I_{i,g+1} &= E\left(-\frac{\partial^2 l}{\partial \pi_i \partial R} \right) = -\frac{2(1 - R)\pi_i^2 m_i}{(R\pi_i^2 - 2\pi_i + 1)(1 - R\pi_i)}, \\
I_{ij} &= E\left(-\frac{\partial^2 l}{\partial \pi_i \partial \pi_j} \right) = 0, \quad i \neq j, \\
I_{g+1,g+1} &= E\left(-\frac{\partial^2 l}{\partial R^2} \right) = \sum_{i=1}^{g} \frac{\pi_i^2 m_i (R\pi_i - 2\pi_i + 1)}{R(R\pi_i^2 - 2\pi_i + 1)(1 - R\pi_i)}.
\end{aligned}$$

The $(g + 1)\times(g + 1)$ dimension information matrix is denoted as $I(\pi_1, \ldots, \pi_g; R)$.

2.2 Model-Based Method

Correlated outcomes are collected in many research areas and occur for various reasons. Valid scientific inferences rely on adequately accounting for the correlation among outcomes within subjects. This type of within-subject correlation may be due to a single outcome repeatedly measured over time on the same subject, as in longitudinal studies, or maybe due to multiple outcomes measured one or more times on the same subject, as in clinical trials involving multiple investigative endpoints. Correlation may also be due to a membership relationship among units (families).

The Generalized Estimating Equations (GEE) approach introduced by Liang and Zeger (1986) is a method for analyzing correlated outcome data, when those data could have been modeled using GLMs if there were no correlated outcomes. By specifying possible working correlation structures to account for the within-subject correlations, this approach estimates model parameters by iteratively solving a system of equations based on quasi-likelihood distributional assumptions. An incorrect specification can affect the efficiency of the parameter estimates. The most commonly used within-subject correlation matrices are:

- Independence: Repeated observations are uncorrelated.
- Unspecified (unstructured): Correlations within any two responses are unknown and need to be estimated.
- Exchangeable: The correlation between any two responses of the ith individual is the same.
- Autoregressive of first order [AR(1)]: Assuming the interval length is the same between any two observations.

GEE models can be constructed and analyzed using SAS. The GENMOD procedure is a powerful tool in SAS for conducting generalized linear regressions and the extension to Generalized Estimating Equations where correlated outcomes must be considered. Our interest is to test whether the disease rates of the g groups are identical and compare Type I error and power of this model-based test and Ma and Wang's Score test. Therefore, we are interested in the relationship between the outcome of eye disease and group. We treat the measurements of two eyes

```
proc genmod data=stacked_data descending;
freq n;
class group id;
model resp=group / link=logit dist=bin;
repeated subject=id / type=UN corrw;
by simcase;
contrast 'group' group 1 -1 0 0, group 0 1 -1 0, group 0 0 1 -1;
run;
```

Figure 1: SAS procedure for analyzing stacked data.

from each subject as repeated observations, the binary outcome of individual eyes as response variables, and the group as a predictor variable. We assign 0 to the eye without disease and 1 as the eye with a disease, and we treat the left eye and right eye with no difference. Moreover, in the GENMOD procedure, we specify "unstructured" as the structure of the working correlation matrix used to model the correlation of the responses from subjects. Data sets, as well as the empirical type I error rates and powers of Ma and Wang's Score test, are simulated using MATLAB.

For example, when $g = 4$, the following is the core code to implement the GEE method's procedure. The original data, generated from MATLAB, are $(\text{n1}, \text{n2}, \ldots, \text{n20})$, corresponding to the 20 cells in Table 1, which is $(m_{01}, m_{11}, m_{21}, m_{02}, m_{12}, m_{22}, m_{03}, m_{13}, m_{23}, m_{04}, m_{14}, m_{24}, n_{01}, n_{11}, n_{02}, n_{12}, n_{03}, n_{13}, n_{04}, n_{14})$. The original dataset is reshaped into a stacked format, where each row represents an individual eye, as prepared using the code provided in Appendix Figure A1. Key variables include id (individual identifier), resp (binary response), eye (Left or Right), group (treatment or exposure group), and n (frequency of observations). Each row is tagged with its simulation case (simcase). The analysis is conducted using the procedure shown in Figure 1. Using PROC GENMOD, the binary response variable (resp) is modeled with a logit link and binomial distribution. The independent variable is group, capturing differences across four groups. The correlation within individuals (id) is handled using an unstructured covariance matrix (type=UN) to model the dependency between paired observations (left and right eyes). The "by simcase" statement ensures that the analysis is repeated separately for each simulation case. The "contrast" statement specifies linear comparisons between group effects, enabling pairwise evaluation of differences.

3 Simulation Studies

We evaluate the testing performances by comparing the GEE method and Ma and Wang's Score test. Compared to Donner's adjusted chi-square method and other alternative testing procedures, Ma and Wang's Score test has satisfactorily type I error control and produces higher power regardless of parameter configurations, and therefore is highly recommended (Ma and Wang, 2021). Thus, the performance of the Score test can be seen as a standard worth comparing with.

3.1 Empirical Type I Error

First, we investigate the behavior of the type I error rates of the two procedures with balanced sample size $m_1 = \cdots = m_g = n_1 = \cdots = n_g = 20, 40$, $\pi_0 = 0.3, 0.5$, number of groups

$g = 2, 4, 8$, and correlation coefficient $\rho = 0.4, 0.5, 0.7$. In each configuration, 10,000 replications are generated based on the null hypothesis that the disease rates throughout g groups are identical. We also consider some cases with unbalanced sample sizes: $(m_1, \ldots, m_g) = (n_1, \ldots, n_g) = (20, 40), (20, 20, 40, 40), (20, 20, 30, 30, 40, 40, 50, 50)$ for $g = 2, 4, 8$, respectively.

For each of the replications, the observed data $\tilde{D}$ are generated from $(m_{0i}, m_{1i}, m_{2i}) \sim \text{Multinomial}(m_i, (R_0\pi_0^2 - 2\pi_0 + 1, 2\pi_0(1 - R_0\pi_0), R_0\pi_0^2))$ and $n_{1i} \sim \text{Binomial}(n_i, \pi_0)$, $i = 1, \ldots, g$, where $R_0 = \frac{(1-\pi_0)\rho_0}{\pi_0} + 1$. Here we assume equal R between two eyes of the same person across groups. We reject the null hypothesis $H_0 : \pi_1 = \cdots = \pi_g$ if the estimated p-value is less than 0.05. The empirical Type I error rate is calculated as the proportion of simulations in which the null hypothesis is incorrectly rejected, determined by dividing the number of rejections by the total number of simulations (10,000).

3.2 Empirical Power

We also evaluate the performance of powers for the two methods. We consider the two alternative hypotheses with

$$H_{1A} : \pi = (0.25, 0.4), (0.25, 0.3, 0.35, 0.4), (0.25, 0.3, 0.35, 0.4, 0.25, 0.3, 0.35, 0.4),$$
$$H_{1B} : \pi = (0.2, 0.4), (0.2, 0.2, 0.4, 0.4), (0.2, 0.2, 0.4, 0.4, 0.2, 0.2, 0.4, 0.4)$$

for $g = 2, 4, 8$, respectively. Correlation coefficient $\rho = 0.4, 0.5, 0.7$. The same settings are applied to both balanced and unbalanced sample designs. For each design, we follow the same process used to compute the empirical Type I error to calculate the power. Under the alternative hypothesis, power is determined as the proportion of simulations in which the null hypothesis is correctly rejected.

3.3 Results

Following Tang et al. (2008), we say a test is liberal if the ratio of its empirical type I error rate to the nominal type I error rate is greater than 1.2 (e.g., empirical type I error rate > 0.06 for $\alpha = 0.05$), conservative if the ratio is less than 0.8 (e.g., empirical type I error rate < 0.04), and robust otherwise.

We denote the T_{SC} as Ma's Score test and T_{RM} as the GEE method. As seen in Table 2, for 36 balanced configurations, both methods provide robust type I error control. The GEE method produced an inflated type I error rate compared to the Score test for most configurations. For unbalanced configurations, see Table 3; similar behavior is shown in the balanced sample size case.

Table 4 shows the powers of the balanced cases, and Table 5 shows the power of the unbalanced cases. In general, test power for H_{1A} is significantly lower than it for H_{1B}. This is because the difference between H_{1A} and the null hypothesis is smaller than the difference between H_{1B} and the null hypothesis. In balanced design, for $g = 2$, the power of T_{RM} is slightly less than T_{SC} regardless of the alternative hypothesis setting. For $g = 4$ or 8, the empirical powers of the two tests do not show a specific pattern, but their results are very close. When the sample size increases from 20 to 40, both of the powers of the two tests improve a lot. Also, for both methods, the power decrease as correlation coefficient ρ increase for almost all configurations. That is, if the correlation of disease between the two eyes is high, we would expect both approaches to have low power. In unbalanced design, for $g = 2$ and 4, the power of T_{RM} is slightly higher than T_{SC} regardless of the alternative hypothesis setting. But for $g = 8$, the power of T_{RM} is slightly lower than T_{SC}.

Table 2: The empirical type I error rates for the case of balanced sample sizes.

m	n	π_0	ρ	$g=2$		$g=4$		$g=8$	
				T_{SC}	T_{RM}	T_{SC}	T_{RM}	T_{SC}	T_{RM}
20	20	0.3	0.4	0.0441	0.0486	0.0403	0.0471	0.0408	0.0560
			0.5	0.0472	0.0521	0.0520	0.0546	0.0497	0.0552
			0.7	0.0435	0.0497	0.0433	0.0496	0.0402	0.0577
		0.5	0.4	0.0504	0.0525	0.0459	0.0510	0.0485	0.0480
			0.5	0.0432	0.0508	0.0418	0.0487	0.0403	0.0512
			0.7	0.0416	0.0561	0.0430	0.0487	0.0413	0.0490
40	40	0.3	0.4	0.0517	0.0524	0.0524	0.0497	0.0480	0.0537
			0.5	0.0560	0.0495	0.0505	0.0521	0.0527	0.0505
			0.7	0.0477	0.0486	0.0504	0.0524	0.0484	0.0480
		0.5	0.4	0.0504	0.0547	0.0460	0.0502	0.0470	0.0516
			0.5	0.0490	0.0518	0.0517	0.0451	0.0465	0.0515
			0.7	0.0512	0.0541	0.0476	0.0478	0.0479	0.0506

Table 3: The empirical type I error rates for the case of unbalanced sample sizes.

π_0	ρ	$g=2$		$g=4$		$g=8$	
		T_{SC}	T_{RM}	T_{SC}	T_{RM}	T_{SC}	T_{RM}
0.3	0.4	0.0451	0.0542	0.0454	0.0494	0.0436	0.0509
	0.5	0.0508	0.0507	0.0468	0.0553	0.0521	0.0526
	0.7	0.0475	0.0485	0.0478	0.0510	0.0434	0.0519
0.5	0.4	0.0497	0.0494	0.0451	0.0485	0.0506	0.0475
	0.5	0.0419	0.0503	0.0465	0.0507	0.0477	0.0501
	0.7	0.0478	0.0537	0.0442	0.0510	0.0442	0.0460

Sample sizes: $(m_1,\ldots,m_g)=(n_1,\ldots,n_g)=(20,40)$, $(20,20,40,40)$, $(20,20,30,30,40,40,50,50)$ for $g=2,4,8$.

Table 4: The empirical powers for the case of balanced sample sizes.

m	n	ρ	H_{1A}						H_{1B}					
			$g=2$		$g=4$		$g=8$		$g=2$		$g=4$		$g=8$	
			T_{SC}	T_{RM}	T_{SC}	T_{RM}	T_{SC}	T_{RM}	T_{SC}	T_{RM}	T_{SC}	T_{RM}	T_{SC}	T_{RM}
20	20	0.4	0.3694	0.3598	0.2592	0.2629	0.3675	0.3783	0.6061	0.5912	0.7720	0.7628	0.9326	0.9344
		0.5	0.3688	0.3575	0.2516	0.2605	0.3632	0.3733	0.5969	0.5943	0.7585	0.7512	0.9311	0.9285
		0.7	0.3626	0.3528	0.2589	0.2649	0.3630	0.3700	0.6004	0.5937	0.7435	0.7441	0.9204	0.9201
40	40	0.4	0.6725	0.6252	0.5474	0.5129	0.7481	0.7063	0.9089	0.8722	0.9861	0.9761	0.9999	0.9994
		0.5	0.6715	0.6272	0.5361	0.5026	0.7493	0.7063	0.9051	0.8752	0.9481	0.9752	0.9998	0.9994
		0.7	0.6653	0.6294	0.5307	0.5104	0.7323	0.6979	0.9037	0.8766	0.9846	0.9743	0.9997	0.9996

Table 5: The empirical powers for the case of unbalanced sample sizes.

ρ	H_{1A}						H_{1B}					
	$g=2$		$g=4$		$g=8$		$g=2$		$g=4$		$g=8$	
	T_{SC}	T_{RM}	T_{SC}	T_{RM}	T_{SC}	T_{RM}	T_{SC}	T_{RM}	T_{SC}	T_{RM}	T_{SC}	T_{RM}
0.4	0.4620	0.4792	0.3398	0.3717	0.6486	0.6395	0.7279	0.7358	0.8787	0.8916	0.9984	0.9975
0.5	0.4605	0.4840	0.3513	0.3826	0.6415	0.6354	0.7267	0.7433	0.8666	0.8872	0.9982	0.9981
0.7	0.4474	0.4769	0.3297	0.3739	0.6289	0.6156	0.7052	0.7348	0.8647	0.8886	0.9976	0.9966

Sample sizes: $(m_1,\ldots,m_g)=(n_1,\ldots,n_g)=(20,40)$, $(20,20,40,40)$, $(20,20,30,30,40,40,50,50)$ for $g=2,4,8$.

4 Conclusions

In this article, we extend the analysis of combined correlated bilateral and unilateral data by applying the GEE model to test the equality of event proportions. This approach allows for greater flexibility and applicability. Specifically, the GEE model accommodates both discrete and continuous explanatory variables, offering a versatile framework for analyzing complex datasets. Furthermore, GEE facilitates the inclusion of additional covariates, making it an invaluable tool for further research in more comprehensive settings.

Our study compares the performance of the GEE approach and the Score test under various scenarios, including balanced and unbalanced sample sizes. Simulation results demonstrate that the GEE model performs comparably to the Score test in most cases. These findings underscore the practical utility of GEE for repeated measures analysis.

However, we also acknowledge the limitations of the GEE approach. The method involves extensive computations and does not provide an explicit form for the test statistics, which could limit its simplicity and broader application. As Ma et al. (2015) noted, the explicit form of the Score test statistic is advantageous for its simplicity and for enabling further developments, such as exact tests.

In summary, Our study demonstrates that, while the GEE method is effective for practical applications, the Score test offers significant advantages in terms of its potential for further refinement and in-depth statistical analysis. Its clear formulation and strong performance make it a crucial tool for advancing the field and improving the accuracy and depth of statistical models.

Supplementary Material

This Supplementary Material contains SAS scripts for analyzing data from two groups ($g=2$) for bilateral and unilateral $m=n=20$; true event proportion $\pi_0=0.5$; Correlation $\rho=0.4$.

Files:

1. 'README.txt': The explanation of the SAS scripts.
2. 'sim02_20_20_0.5_0.4.sas': Prepares the dataset for analysis.
3. 'analysis.sas': Performs statistical analysis.

Appendix

```
data stacked_data;
set original_data;
id=1;n=n1; resp=0; eye='L'; group=1; simcase=_N_;output;
id=1;n=n1; resp=0; eye='R'; group=1; simcase=_N_;output;
id=2;n=n2; resp=0; eye='L'; group=1; simcase=_N_;output;
id=2;n=n2; resp=1; eye='R'; group=1; simcase=_N_;output;
id=3;n=n3; resp=1; eye='L'; group=1; simcase=_N_;output;
id=3;n=n3; resp=1; eye='R'; group=1; simcase=_N_;output;
id=4;n=n4; resp=0; eye='L'; group=2; simcase=_N_;output;
id=4;n=n4; resp=0; eye='R'; group=2; simcase=_N_;output;
id=5;n=n5; resp=0; eye='L'; group=2; simcase=_N_;output;
id=5;n=n5; resp=1; eye='R'; group=2; simcase=_N_;output;
id=6;n=n6; resp=1; eye='L'; group=2; simcase=_N_;output;
id=6;n=n6; resp=1; eye='R'; group=2; simcase=_N_;output;
id=7;n=n7; resp=0; eye='L'; group=3; simcase=_N_;output;
id=7;n=n7; resp=0; eye='R'; group=3; simcase=_N_;output;
id=8;n=n8; resp=0; eye='L'; group=3; simcase=_N_;output;
id=8;n=n8; resp=1; eye='R'; group=3; simcase=_N_;output;
id=9;n=n9; resp=1; eye='L'; group=3; simcase=_N_;output;
id=9;n=n9; resp=1; eye='R'; group=3; simcase=_N_;output;
id=10;n=n10; resp=0; eye='L'; group=4; simcase=_N_;output;
id=10;n=n10; resp=0; eye='R'; group=4; simcase=_N_;output;
id=11;n=n11; resp=0; eye='L'; group=4; simcase=_N_;output;
id=11;n=n11; resp=1; eye='R'; group=4; simcase=_N_;output;
id=12;n=n12; resp=1; eye='L'; group=4; simcase=_N_;output;
id=12;n=n12; resp=1; eye='R'; group=4; simcase=_N_;output;
id=13;n=n13; resp=0; eye='L'; group=1; simcase=_N_;output;
id=14;n=n14; resp=1; eye='R'; group=1; simcase=_N_;output;
id=15;n=n15; resp=0; eye='L'; group=2; simcase=_N_;output;
id=16;n=n16; resp=1; eye='R'; group=2; simcase=_N_;output;
id=17;n=n17; resp=0; eye='L'; group=3; simcase=_N_;output;
id=18;n=n18; resp=1; eye='R'; group=3; simcase=_N_;output;
id=19;n=n19; resp=0; eye='L'; group=4; simcase=_N_;output;
id=20;n=n20; resp=1; eye='R'; group=4; simcase=_N_;output;
keep id n resp eye group simcase;
run;
```

Figure A1: SAS code for data preparation.

References

Dallal GE (1988). Paired Bernoulli trials. *Biometrics*, 44(1): 253–257. https://doi.org/10.2307/2531913

Donner A (1989). Statistical methods in ophthalmology: An adjusted chi-square approach. *Biometrics*, 45(2): 605–611. https://doi.org/10.2307/2531501

Liang K, Zeger SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1): 13–22. https://doi.org/10.1093/biomet/73.1.13

Lu W, Ning R, Diao K, Ding Y, Chen R, Zhou L, et al. (2022). Comparison of two main orthokeratology lens designs in efficacy and safety for myopia control. *Frontiers in Medicine*, 9: 798314. https://doi.org/10.3389/fmed.2022.798314

Ma C, Shan G, Liu S (2015). Homogeneity test for correlated binary data. *PLoS ONE*, 10(4): e0124337. https://doi.org/10.1371/journal.pone.0124337

Ma C, Wang K (2021). Testing the homogeneity of proportions for combined unilateral and bilateral data. *Journal of Biopharmaceutical Statistics*, 31(5): 686–704. https://doi.org/10.1080/10543406.2021.1946693

Mandel EM, Bluestone CD, Rockette HE, Blatter MM, Reisinger KS, Wucher FP, et al. (1982). Duration of effusion after antibiotic treatment for acute otitis media: Comparison of cefaclor and amoxicillin. *The Pediatric Infectious Disease Journal*, 1(5): 310–316. https://doi.org/10.1097/00006454-198209000-00006

Pei Y, Tang M, Guo J (2008). Testing the equality of two proportions for combined unilateral and bilateral data. *Communications in Statistics. Simulation and Computation*, 37(8): 1515–1529. https://doi.org/10.1080/03610910802140232

Rosner B (1982). Statistical methods in ophthalmology: An adjustment for the intraclass correlation between eyes. *Biometrics*, 38(1): 105–114. https://doi.org/10.2307/2530293

Tang N, Tang M, Qiu S (2008). Testing the equality of proportions for correlated otolaryngologic data. *Computational Statistics & Data Analysis*, 52(7): 3719–3729. https://doi.org/10.1016/j.csda.2007.12.017