Statistics Innovation

ARTICLE

Open Access

https://doi.org/10.48130/stati-0026-0003

Statistics Innovation **2026**, 3: e003

Homogeneity test of proportions for combined unilateral and bilateral data via GEE and MLE approaches

Jia Zhou and Chang-Xing Ma*

Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, USA

* Correspondence: cxma@buffalo.edu (Ma CX)

Abstract

In clinical trials involving paired organs such as eyes, ears, and kidneys, binary outcomes may be collected bilaterally or unilaterally. In such combined datasets, bilateral outcomes exhibit intra-subject correlation, while unilateral outcomes are assumed independent. The generalized Estimating Equations (GEE) approach for testing homogeneity of proportions across multiple groups for combined unilateral and bilateral data were investigated, and compared with three likelihood-based statistics (likelihood ratio, Wald-type, and score) under Rosner's constant R model, and Donner's equal correlation ρ model. Monte Carlo simulations evaluate empirical type I error and power under varied sample sizes and parameter settings. The GEE and score tests show superior type I error control, outperforming likelihood ratio and Wald-type tests. Applications to two real datasets in otolaryngologic and ophthalmologic studies illustrate the methods. We recommend the GEE and score tests for homogeneity testing and suggest GEE for more complex models with covariates, while favoring the score statistic for small-sample exact tests due to its computational efficiency.

Citation: Zhou J, Ma CX. 2026. Homogeneity test of proportions for combined unilateral and bilateral data via GEE and MLE approaches. *Statistics Innovation* 3: e003 https://doi.org/10.48130/stati-0026-0003

Introduction

In comparative clinical trials, binary bilateral observations often consist of paired organ (e.g., eyes, ears, and kidneys) records in dichotomous form, with '1' denoting a cured or affected status, and '0' otherwise. In practice, such datasets often comprise a mixture of bilateral and unilateral observations, since some subjects contribute paired organ records, while others only a single record, for reasons such as prior surgical removal, congenital absence, or missing measurement for one organ. It is therefore essential to analyze the combined bilateral and unilateral data together, rather than discarding unilateral observations, to avoid loss of valuable information and potential bias in inference. Furthermore, appropriate modeling of the intra-subject correlation inherent in paired bilateral outcomes is necessary to accurately capture the dependence structure and ensure valid statistical conclusions. Rosner[1] first proposed the 'constant R model' to test whether the proportion of affected eyes is equal across all groups, assuming equal dependence between the two eyes of the same individual. Dallal[2] critiqued Rosner's model for its poor fit when the binary status (e.g., cured or affected) is present bilaterally, yet shows substantially different prevalences across groups. To address this, he proposed a model where the conditional probability of a response in one organ, given a response in the other, is a constant. Donner[3] introduced an alternative approach assuming a constant intra-subject correlation across all individuals, which Thompson[4] subsequently demonstrated to be robust through simulation studies.

Recent literature (2015–2025) demonstrates that these three models remain central to modern analyses of bilateral and combined uniliateral-bilateral data. Independent research groups continue to apply Rosner's, Dallal's, and Donner's methods to a variety of clinical contexts, including ophthalmology, otolaryngology, and other paired organ investigations. Several recent studies employ Rosner's model for the analysis of correlated bilateral outcomes[5–11], while others rely on Dallal's model in applications involving correlated bilateral data[12–16], or Donner's constant correlation framework to evaluate dependence in bilateral measurements[17–20]. These works collectively reaffirm that bilateral and combined unilateral-bilateral data continue to represent an active statistical research area, and that classical dependence models are still widely used in modern biomedical statistics.

Testing the homogeneity of proportions addresses a fundamental inferential question in comparative studies, namely whether multiple groups share a common marginal probability of response while accounting for within-subject dependence in correlated outcomes. Methods based on maximum likelihood estimates (MLEs), and likelihood-based tests under the foregoing models have been developed in numerous subsequent studies. For example, Tang et al.[21] proposed an asymptotic method for testing the equality of proportions between two groups under Rosner's model for correlated binary data. Ma et al.[5], and Ma & Liu[17] investigated three tests (likelihood ratio, Wald-type, and score) for testing homogeneity among $g \geq 2$ groups under Rosner's and Donner's models, respectively. Ma & Wang K.[7], and Ma & Wang H.[19] further examined several test procedures to test the homogeneity of general $g \geq 2$ proportions for combined unilateral and bilateral data under Rosner's and Donner's models, respectively.

Alternatively, regression models for correlated paired organ data can be used to perform hypothesis testing[22,23], with the generalized linear mixed model (GLMM)[24–27], and generalized estimating equations (GEE)[28–30] being the most common approaches. Both methods are implemented in standard statistical software such as SAS, R, and Stata. The GEE method, introduced by Liang & Zeger for longitudinal data[28–30], and later extended to clustered data[31], generalizes the generalized liner model (GLM) framework to account for within-cluster dependence via estimating equations. It has been widely adopted for longitudinal and clustered data, and is often preferred over GLMM because it yields consistent parameter estimates and robust (co)variance estimates even when the 'working' correlation structure is misspecified[32–35].

Bilateral observations from the same subject can be regarded as repeated measurements on that individual, making the GEE framework a natural choice for analyzing such correlated binary data. In this paper, the GEE method is employed to analyze combined unilateral and bilateral data, and compare its performance with established likelihood-based procedures, including the likelihood ratio, Wald-type, and score tests. Specifically, the homogeneity of proportions for combined unilateral and bilateral data under both Rosner's and Donner's models are investigated. A recent study by Zhang & Ma[36] compared the score test with the GEE method for the equal proportion test in the combined data framework under Rosner's model. The present work extends this line of research by conducting a systematic comparison between the GEE approach, and the three likelihood-based methods under both Rosner's and Donner's models, providing a broader evaluation in this context.

The rest of the paper is organized as follows. The Methods section presents the MLE approaches using likelihood ratio, Wald-type and score statistics, as well as the GEE method for correlated binary outcomes and its implementation in SAS. The Results section reports the numerical results, including a simulation study evaluating the empirical type I error rates and powers of these methods under different models, and two real-world applications in otolaryngologic and ophthalmologic studies. The paper concludes with a discussion of the findings and their implications.

Methods

Consider a study involving the combined unilateral and bilateral date, where in the i-th group ($i = 1, \ldots, g$), there are m_{+i} subjects who contribute data from both paired organs (bilateral), and n_{+i} subjects who contribute data from one of the paired organs (unilateral), respectively. Let m_{ri} be the number of bilateral subjects who have r ($r = 0, 1, 2$) organs cured or affected, and n_{ri} be the number of unilateral subjects who have r ($r = 0, 1$) organs cured or affected, such that:

$$m_{ri} = \sum_{j=1}^{m_{+i}} I\left(Z_{ij1} + Z_{ij2} = r\right), \quad r = 0, 1, 2$$

$$n_{ri} = \sum_{j=1}^{n_{+i}} I\left(Z_{ijk} = r\right), \quad r = 0, 1;\ k = 1 \text{ or } 2$$

where, Z_{ijk} denotes the response (1 cured or affected; 0 otherwise) of the k-th paired organ ($k = 1, 2$) of the j-th subject in the i-th group. The data structure is summarized in Table 1, where a subscript '+' indicates the summation over the corresponding index.

The proportion of cured or affected organs in the i-th group is assumed to be $Pr\ (Z_{ijk} = 1) = \pi_i$. Given m_{+i} and n_{+i} in the i-th group, (m_{0i}, m_{1i}, m_{2i}) follows trinomial distribution and (n_{0i}, n_{1i}) follows binomial distribution, i.e.,

$$(m_{0i}, m_{1i}, m_{2i}) \sim Trinomial(m_{+i}, p_{0i}, p_{1i}, p_{2i}), \quad n_{1i} \sim Binomial(n_{+i}, \pi_i) \quad (1)$$

where, the joint probabilities p_{ri}'s ($r = 0, 1, 2$) read

$$\begin{aligned} p_{2i} &= Pr\left(Z_{ij1} = 1, Z_{ij2} = 1\right) = E\left(Z_{ij1}Z_{ij2}\right) = \pi_i\left[\pi_i + (1-\pi_i)\,\mathrm{Corr}\left(Z_{ij1}, Z_{ij2}\right)\right] \\ p_{1i} &= \sum_{k=1}^{2} Pr\left(Z_{ijk} = 1, Z_{ij,3-k} = 0\right) = 2(\pi_i - p_{2i}) = \\ &\quad 2\pi_i(1-\pi_i)\left[1 - \mathrm{Corr}\left(Z_{ij1}, Z_{ij2}\right)\right] \\ p_{0i} &= Pr\left(Z_{ij1} = 0, Z_{ij2} = 0\right) = 1 - p_{1i} - p_{2i} = \\ &\quad (1-\pi_i)\left[1 - \pi_i + \pi_i\,\mathrm{Corr}\left(Z_{ij1}, Z_{ij2}\right)\right] \end{aligned} \quad (2)$$

with Corr (Z_{ij1}, Z_{ij2}) being the intra-subject correlation between the two responses from the j-th subject in the i-th group.

In what follows, two parametric models proposed by Rosner[1] and Donner[3] are considered to address the intra-subject correlation for the binary data. Under Rosner's model, the conditional probability is specified as $Pr\ (Z_{ijk} = 1 \mid Z_{ij,3-k} = 1) = R\pi_i$, where R is a scalar parameter. Based on this specification, the intra-subject correlation becomes Corr (Z_{ij1}, Z_{ij2}) = $(R - 1)\pi_i/(1 - \pi_i)$. Under Donner's model, a constant correlation ρ is assumed across all g groups, i.e., Corr (Z_{ij1}, Z_{ij2}) = ρ. Under either model, the joint probabilities in Eq. (2) can be written in terms of π_i and the nuisance parameter κ ($\kappa = R$ under Rosner's and $\kappa = \rho$ under Donner's model).

Table 1. Frequency table for the number of cured or affected organs for subjects in g groups.

No. of cured or affected organs	Group				Total
	1	2	…	g	
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_{++}

MLE approach

Let $\beta = (\pi_1, \ldots, \pi_g, \kappa)^T$ be the vector of parameters. For given observation:

$$(m, n) = \left(m_{01}, m_{11}, m_{2,1}, \cdots, m_{0g}, m_{1g}, m_{2g}, n_{01}, n_{11}, \cdots, n_{0g}, n_{1g}\right)$$

the log-likelihood function reads:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{g}\sum_{r=0}^{2} m_{ri}\log(p_{ri}) + \sum_{i=1}^{g}\left[n_{0i}\log(1-\pi_i) + n_{1i}\log(\pi_i)\right] + \text{const} \quad (3)$$

where, the term 'const' denotes a constant depending on (m, n).

Our interest is the homogeneity test of proportions across the g groups. Thus, the hypotheses are:

$$H_0:\ \pi_1 = \cdots = \pi_g = \pi, \quad \text{versus} \quad H_1:\ \text{some of } \pi_i\text{'s are not equal.} \quad (4)$$

The MLEs $\hat{\boldsymbol{\beta}}_0$ under H_0 can be solved analytically, while the MLEs $\hat{\boldsymbol{\beta}}_1$ under H_1 have no closed-form solutions and must be computed using an iterative method. Detailed procedures for obtaining $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ can be found in the works of Ma & Wang K.[7] and Ma & Wang H.[19], corresponding to Rosner's and Donner's models, respectively. Based on these MLEs, three likelihood-based test statistics are considered and defined as follows:

Likelihood Ratio (LR) test:

$$Q_{LR} = 2\left[l\left(\hat{\boldsymbol{\beta}}_1\right) - l\left(\hat{\boldsymbol{\beta}}_0\right)\right] \quad (5)$$

Wald-type test:

$$Q_W = \left(\boldsymbol{\beta}^T\boldsymbol{C}\right)\left[\boldsymbol{C}^T\boldsymbol{I}^{-1}(\boldsymbol{\beta})\boldsymbol{C}\right]^{-1}\left(\boldsymbol{C}^T\boldsymbol{\beta}\right)\Big|_{\beta=\hat{\beta}_0} \quad (6)$$

where, C^T is a $(g - 1) \times (g + 1)$ hypothesis matrix with $(C^T)_{ii} = 1$, $(C^T)_{i,i+1} = -1$ for $i = 1, \ldots, g - 1$, and $(C^T)_{ij} = 0$ otherwise; $I^{-1}(\beta)$ is the inverse of the $(g + 1) \times (g + 1)$ Fisher's information matrix whose explicit forms under Rosner's and Donner's models are given in Appendices in the works of Ma & Wang K.[7], and Ma & Wang H.[19], respectively.

Score test:

$$Q_S = \boldsymbol{U}\boldsymbol{I}^{-1}(\boldsymbol{\beta})\boldsymbol{U}^T\big|_{\beta=\hat{\beta}} \quad (7)$$

where, $U = (\partial l/\partial\pi_1, \ldots, \partial l/\partial\pi_g, 0)$ is the score function, and $I^{-1}(\beta)$ is the inverse of Fisher's information matrix $I(\beta)$.

It can be shown that the above three test statistics asymptotically follow the chi-square distribution with degrees of freedom $df = g - 1$ under H_0, i.e., $Q_{LR}, Q_W, Q_S \xrightarrow{d} \chi^2_{g-1}$.

GEE approach

Treating the bilateral observations as repeated measurements on the same subject, the GEE method can be employed to perform an equal proportion test with the mean model $\mathbb{E}(Z)_{ij} = \mu_{ij} = \pi_i \mathbb{1}_{k_{ij}}$ ($\mathbb{1}_{k_{ij}}$ denotes a $k_{ij} \times 1$ column vector of ones), and the estimating equations $S(\beta) = \sum_{j=1}^{n_i} \partial\mu_{ij}^T/\partial\pi_i V_{ij}^{-1}(Z_{ij} - \mu_{ij}) = 0$, where Z_{ij} denotes the response(s) of the j-th subject in the i-th group, which is a 2×1 vector for bilateral data and a scalar for unilateral data, $k_{ij} = 2$ for bilateral data and $k_{ij} = 1$ for unilateral data, and $V_{ij} = A_{ij}^{1/2} R_{ij} A_{ij}^{1/2}$ is the "working covariance" matrix. Here, R_{ij} is the "working correlation" and A_{ij} is the diagonal matrix of marginal variances of Z_{ij}. The working correlation matrix R_{ij} generally depends on unknown parameters that must be estimated. In the case of paired organ data, R_{ij} contains only a single parameter. Various working correlation structures can be specified, such as independent, exchangeable, and unstructured correlations. Estimates for the parameters of both the mean model and the working covariance are obtained through an iterative procedure based on the estimating equations (see, e.g., the GENMOD procedure in the SAS user's guide[37]).

Several statistical software systems can implement the GEE approach for analyzing combined unilateral and bilateral data. In this study, the SAS GENMOD procedure is used. Table 2 illustrates the structure of the combined unilateral and bilateral data for the case $g = 2$ in a single simulation.

An example of SAS code invoking the GENMOD procedure for $g = 2$ is presented in Table 3. In this example, the link function is the identity link specified with link = identity, and the unstructured working correlation matrix is specified with type = un. In addition, the by replicate statement repeats the analysis for each simulation case, and the contrast statement enables pairwise evaluation of differences, corresponding to the equal proportion test. Note that with repeated statement, the default statistic provided by the contrast statement is a generalized score statistic Q_{GS} computed based on the generalized score function $S(\beta)$[37]. A generalized Wald-type statistic can be obtained with the wald option in the contrast statement. However, the generalized Wald-type statistic is known to exhibit poorer finite-sample performance and less accurate control of nominal significance levels compared with the generalized score statistic[37,38].

Table 2. The structure of stacked data ($g = 2$) in the first simulation (replicate = 1).

Sub_id	Response	Group	Count	Replicate
1	0	1	m_{01}	1
1	0	1	m_{01}	1
2	1	1	m_{11}	1
2	0	1	m_{11}	1
3	1	1	m_{21}	1
3	1	1	m_{21}	1
4	0	2	m_{02}	1
4	0	2	m_{02}	1
5	1	2	m_{12}	1
5	0	2	m_{12}	1
6	1	2	m_{22}	1
6	1	2	m_{22}	1
7	0	1	n_{01}	1
8	1	1	n_{11}	1
9	0	2	n_{02}	1
10	1	2	n_{12}	1

Table 3. Pseudo SAS code using PROC GENMOD.

```
* stacked data (g = 2) file w/ name <data_g2_stacked>;
* var names same with those in the above table (Table 2);
proc genmod data = data_g2_stacked descending;
freq count;
class group sub_id;
model response = group/link = identity dist = bin;
repeated subject = sub_id/type = un corrw;
by replicate;
contrast 'group' group 1 -1;
run;
```

Results

Simulation study

A simulation study was performed to assess the performance of the three likelihood-based test statistics along with the GEE-based generalized score test, by investigating the empirical type I error and powers, respectively, under Rosner's and Donner's model. Equal and unequal sample size for $g = 2, 4, 8$ are considered. Specifically, for equal sample size design, we set:

$$m_{+1} = \cdots = m_{+g} = n_{+1} = \cdots = n_{+g} = 20,\ 40$$

and for unequal sample size design, we set:

$$\left(m_{+1}, \cdots, m_{+g}\right) = \left(n_{+1}, \cdots, n_{+g}\right) = (20,\ 40),\ (20,\ 20,\ 40,\ 40),\ (20,\ 20,\ 30,\ 30,\ 40,\ 40,\ 50,\ 50)$$

Empirical type I error

To compute the empirical type I error rates, the dataset is generated according to the distributions in Eq. (1) with the above sample size designs under $H_0 : \pi_1 = \ldots = \pi_g = \pi_0$ with a pre-specified $\pi_0 = 0.2$, 0.5. Additionally, the intra-subject correlation is set to be $\rho_0 = 0.1$, 0.5, 0.7. Thus, in data generation, $R = R_0 = (1 - \pi_0)\rho_0/\pi_0 + 1$ is set under Rosner's model, and $\rho = \rho_0$ under Donner's model.

After generating the dataset with each sample size design and parameter configuration, the three likelihood-based test statistics Q_{LR}, Q_W, Q_S in Eqs. (5)–(7), and the GEE-based generalized score test statistic denoted by Q_{GS} are computed accordingly. The null hypothesis $H_0 : \pi_1 = \ldots = \pi_g$ is rejected if Test Statistic $> \chi^2_{1-\alpha, g-1}$, where $\chi^2_{1-\alpha, g-1}$ is the quantile function of χ^2_{g-1} valued at $1 - \alpha$. The above simulation is replicated for $N = 100{,}000$ times. Then the empirical type I error rate is calculated as:

$$\widehat{\mathrm{TIE}} = \frac{\sum_{k=1}^{N} I\left(Q_i^{H_0} > \chi^2_{1-\alpha, g-1}\right)}{N} \quad (8)$$

where, $Q_i^{H_0}$ denotes the test statistic resulting from the dataset generated under H_0, and the subscript 'i' represents the type of tests for $i = LR, W, S, GS$.

The results for the empirical type I error with each sample size design and parameter configuration are presented in Table 4 under Rosner's model, and in Table 5 under Donner's model, respectively, where the liberal results are highlighted in boldface (The results are

classified as *liberal* if $\widehat{\text{TIE}} > 0.06$, as *robust* if $0.04 \leq \widehat{\text{TIE}} \leq 0.06$, and as *conservative* if $\widehat{\text{TIE}} < 0.04$). For simplicity, the designs of the sample size $(m_{+1}, \ldots, m_{+g})$, and $(n_{+1}, \ldots, n_{+g})$ are referred to as (m, n) in Tables 4 and 5. The two equal sample size designs, where $m_{+1} = \ldots = m_{+g} = n_{+1} = \ldots = n_{+g} = 20, 40$, are denoted as E_1 and E_2, respectively. The unequal sample size designs are labeled as U_g for g = 2, 4, 8. For instance, U_2 refers to the design where $(m_{+1}, m_{+2}) = (n_{+1}, n_{+2}) = (20, 40)$. Note that these notations also apply to the tables presented for power calculation.

Under Rosner's model, as shown in Table 4, when $g = 2$, the LR and Wald-type tests produce a few liberal results for the sample size design E_1, E_2 and E_1, U_2, respectively. As $g = 4$, the number of liberal results from the Wald-type tests grows substantially across all sample size designs (E_1, E_2 and U_4), whereas the LR tests show a similar number of the liberal results as in the $g = 2$ case. When $g = 8$, the majority of empirical type I error rates from the Wald-type tests are liberal, with several being considerably inflated (rate > 0.10) across all sample size designs. Overall, as the number of groups increases, the LR tests become mildly liberal, while the Wald-type tests exhibit a dramatic increase in both the frequency and amplitude of the liberal behavior. In addition, no clear association is observed between the empirical type I error rates, and either the null proportion π_0 or the intra-subject correlation ρ_0 for the LR or Wald-type tests. In contrast, the score tests consistently yield non-liberal empirical type I error rates across all simulation settings. The GEE tests show robust performance for all sample size designs when $g = 2$ and $g = 4$, and only a few mildly liberal results when $g = 8$. Overall, the results from the score tests closely align with those from the GEE approach and remain closer to the nominal level, as compared to the LR and Wald-type tests.

Under Donner's model, as shown in Table 5, the Wald-type tests begin to produce liberal results when $g = 4$ for sample size designs

Table 4. The empirical type I error rates (in %) under Rosner's model for different test procedures, under $H_0 : \pi_1 = \ldots = \pi_g = \pi_0$ at the nominal level of $\alpha = 0.05$.

(m, n)	π_0	ρ	$g = 2$				$g = 4$				$g = 8$			
			Q_{LR}	Q_W	Q_S	Q_{GS}	Q_{LR}	Q_W	Q_S	Q_{GS}	Q_{LR}	Q_W	Q_S	Q_{GS}
E_1	0.2	0.1	4.41	4.46	4.03	5.15	3.71	4.52	3.43	5.43	2.92	4.63	2.83	5.92
		0.5	5.35	4.44	4.04	5.07	4.67	5.67	3.62	5.68	4.31	**8.25**	3.42	**6.70**
		0.7	5.58	3.35	3.88	5.05	4.03	4.15	3.96	5.86	3.39	5.47	3.15	**6.78**
	0.5	0.1	5.79	**6.50**	5.09	5.10	5.70	**7.48**	4.89	4.90	5.75	**8.34**	4.90	4.91
		0.5	**6.19**	5.79	5.17	5.18	**6.13**	**7.74**	4.91	4.96	**6.33**	**11.34**	4.80	4.90
		0.7	5.18	3.58	5.40	5.03	4.94	4.86	4.30	4.99	4.80	**6.81**	4.02	4.88
E_2	0.2	0.1	4.98	5.18	4.66	5.08	4.62	5.37	4.29	5.24	4.47	5.75	4.21	5.54
		0.5	**6.10**	5.47	5.04	4.96	**6.19**	**7.22**	4.82	5.43	**6.32**	**10.08**	4.75	5.81
		0.7	**6.01**	4.12	5.01	5.03	**6.15**	**6.04**	4.70	5.46	**6.46**	**9.74**	4.70	5.91
	0.5	0.1	5.37	5.80	5.02	4.97	5.42	**6.20**	5.03	5.02	5.39	**6.50**	4.99	4.84
		0.5	5.63	5.38	5.17	5.03	5.64	**6.36**	5.03	4.98	5.66	**7.87**	4.82	4.92
		0.7	5.72	4.64	5.04	5.02	5.91	**6.68**	4.94	4.97	**6.08**	**10.34**	4.89	4.98
U_g	0.2	0.1	4.46	4.79	4.13	5.22	4.10	5.08	3.86	5.56	4.00	5.63	3.84	5.90
		0.5	5.49	5.18	4.41	5.29	5.16	**6.83**	4.16	5.85	5.63	**10.54**	4.40	**6.33**
		0.7	5.21	3.83	5.04	5.34	4.88	5.05	4.18	5.91	5.18	**8.19**	4.18	**6.48**
	0.5	0.1	5.58	**6.33**	5.07	5.10	5.57	**6.94**	5.01	5.01	5.28	**6.83**	4.79	4.75
		0.5	5.79	**6.13**	5.12	4.99	5.88	**7.94**	4.94	4.94	5.99	**9.87**	5.01	5.00
		0.7	5.27	4.77	4.66	5.10	5.17	**6.26**	4.57	4.96	5.36	**10.40**	4.56	5.01

Table 5. The empirical type I error rates (in %) under Donner's model for different test procedures, under $H_0 : \pi_1 = \ldots = \pi_g = \pi_0$ at the nominal level $\alpha = 0.05$.

(m, n)	π_0	ρ	$g = 2$				$g = 4$				$g = 8$			
			Q_{LR}	Q_W	Q_S	Q_{GS}	Q_{LR}	Q_W	Q_S	Q_{GS}	Q_{LR}	Q_W	Q_S	Q_{GS}
E_1	0.2	0.1	**9.53**	4.20	4.24	5.12	4.49	4.70	3.65	5.43	3.48	5.29	3.24	5.91
		0.5	5.12	4.96	4.63	5.05	4.46	5.84	4.16	5.66	4.40	**7.26**	4.09	**6.75**
		0.7	5.00	5.01	4.61	5.01	4.79	**6.10**	4.36	5.86	4.30	**7.23**	3.90	**6.71**
	0.5	0.1	5.43	5.96	5.22	5.21	5.41	**6.68**	5.03	4.99	5.32	**7.39**	4.75	4.75
		0.5	5.35	5.87	5.18	5.17	5.38	**6.73**	5.04	5.00	5.44	**7.78**	4.96	4.90
		0.7	5.36	5.82	5.18	5.17	5.37	**6.75**	5.03	5.01	5.37	**7.74**	4.90	4.91
E_2	0.2	0.1	5.27	4.85	4.61	4.98	4.70	5.30	4.48	5.17	4.65	5.91	4.41	5.51
		0.5	5.13	5.24	5.00	5.05	5.19	**6.03**	4.94	5.41	5.14	**6.82**	4.80	5.69
		0.7	5.00	5.10	4.87	4.88	5.20	5.98	4.98	5.30	5.19	**7.05**	4.81	5.85
	0.5	0.1	5.10	5.38	4.98	4.99	5.17	5.76	4.98	4.96	5.14	**6.11**	4.86	4.84
		0.5	5.03	5.29	4.95	4.96	5.27	5.92	5.12	5.09	5.16	**6.26**	4.91	4.92
		0.7	5.17	5.39	5.11	5.10	5.19	5.88	5.06	5.07	5.17	**6.33**	4.93	4.92
U_g	0.2	0.1	**6.04**	4.69	4.44	5.24	4.22	4.98	4.04	5.38	4.18	5.83	4.07	5.83
		0.5	5.00	5.38	4.76	5.29	4.88	**6.24**	4.62	5.88	4.96	**7.21**	4.60	**6.35**
		0.7	4.98	5.33	4.73	5.34	4.78	**6.20**	4.44	5.76	4.95	**7.36**	4.70	**6.57**
	0.5	0.1	5.22	5.66	5.04	5.04	5.14	**6.11**	4.89	4.83	5.27	**6.51**	4.95	4.91
		0.5	5.16	5.66	5.02	5.00	5.24	**6.31**	4.96	4.98	5.25	**6.73**	4.93	4.92
		0.7	5.15	5.64	5.04	5.03	5.11	**6.25**	4.88	4.85	5.30	**6.83**	5.01	4.99

E_1, E_2, and U_4. As $g = 8$, most empirical type I error rates from the Wald-type tests are liberal across all designs. Compared with the results under Rosner's model, however, the liberal behavior under Donner's model is less severe, with empirical rates generally remaining below approximately 0.08. In contrast, the LR tests yield only a couple of liberal results as $g = 2$ for sample size design E_1 and U_2, when both π_0 and ρ_0 are small. The Score and GEE tests exhibit behavior similar to that observed under Rosner's model.

It should be noted that when the sample size is small (e.g., $(m, n) = E_1$) and both π_0 and ρ_0 are low, some score tests tend to be conservative, whereas the GEE tests may exhibit mild liberal behavior, particularly under Rosner's model. These deviations are likely attributable to sparse cell counts in the resulting contingency tables, which can undermine the validity of asymptotic approximations. As the sample size increases (e.g., $(m, n) = E_2$), however, both the score and GEE tests exhibit improved and stable type I error control.

Powers

The powers are calculated in a similar way as the empirical type I error. The difference is that instead of generating the dataset under H_0, they are generated under certain alternative hypothesis. There are two alternatives, H_{1A} and H_{1B} used for power calculation, which are shown below.

$$H_{1A}: \left(\pi_1, \cdots, \pi_g\right) = (0.25, 0.4), (0.25, 0.3, 0.35, 0.4), (0.25, 0.3, 0.35, 0.4, 0.25, 0.3, 0.35, 0.4)$$

$$H_{1B}: \left(\pi_1, \cdots, \pi_g\right) = (0.2, 0.4), (0.2, 0.2, 0.4, 0.4), (0.2, 0.2, 0.4, 0.4, 0.2, 0.2, 0.4, 0.4)$$

for $g = 2, 4, 8$. Tables 6 and 7 present the results for powers under under H_{1A} and H_{1B} within Rosner's (with $R = 1.2, 1.5, 1.8$), and Donner's (with $\rho = 0.1, 0.5, 0.7$) model, respectively.

It can be seen that the powers under H_{1A} are generally lower than those under H_{1B}. This is because the discrepancy in proportions specified under H_{1A} is smaller compared to that under H_{1B}. Within each power table, the Wald and LR tests are generally the most and second most powerful tests, respectively, for all the sample size designs and parameter configurations. The score tests and GEE tests are less powerful, but their associated powers are comparable to those of Wald and LR tests. In addition, for each given number of groups g and sample design, the powers obtained by the GEE tests decrease as the R (under Rosner's model) or ρ (under Donner's model) increases. The same pattern is observed in the other three likelihood-based tests under Donner's model.

Therefore, based on the simulation results, we conclude that the likelihood-based score test and the GEE test are robust under both Rosner's and Donner's models, with their resulting type I error rates under control and powers that are generally expected, though slightly underestimated. This echoes the findings for the score test in the work of Ma & Wang K.[7] and Ma & Wang H.[19], and in the meantime, recommends the GEE as an alternative approach.

It should be noted that the similar empirical performance observed between the likelihood-based score test Q_S and the GEE-based generalized score test Q_{GS} does not imply that the MLE and GEE approaches are theoretically equivalent. Rather, this similarity arises because both procedures are score-type tests targeting the same null hypothesis and are asymptotically equivalent under correctly specified marginal mean models. In the simulation settings considered here, the estimating equations underlying the GEE generalized score test closely resemble the score equations derived from the likelihood under the assumed dependence structures, leading to comparable finite-sample behavior. Nevertheless, important distinctions remain between the two approaches. The MLE-based score test relies on a fully specified likelihood and explicit modeling assumptions for the within-subject dependence, whereas the GEE-based generalized score test is quasi-likelihood-based and retains validity even when the working correlation structure is misspecified. Therefore, the observed numerical similarity in these simulations should be interpreted as a consequence of the specific study design rather than as evidence of general equivalence between GEE and likelihood-based methods.

Real world example

Two real-world examples are studied to illustrate the equal proportion test. Since there is more than one model available for analysis, a model selection process was first conducted to identify a more suitable model to describe the dataset, following the goodness-of-fit test procedure[39].

Table 6. The powers (in %) under Rosner's model at the nominal level of $\alpha = 0.05$.

(m, n)		R	$g = 2$				$g = 4$				$g = 8$			
			Q_{LR}	Q_W	Q_S	Q_{GS}	Q_{LR}	Q_W	Q_S	Q_{GS}	Q_{LR}	Q_W	Q_S	Q_{GS}
Under H_{1A}	E_1	1.2	40.08	41.85	37.77	39.69	28.79	32.45	26.55	28.68	39.23	45.40	36.65	39.53
		1.5	39.38	40.72	36.03	36.96	28.98	32.62	25.86	26.61	39.36	46.20	35.67	36.40
		1.8	42.22	41.79	37.90	34.83	31.91	34.64	27.64	25.12	44.06	51.39	38.91	34.40
	E_2	1.2	68.10	69.03	66.88	67.45	56.03	58.12	54.52	54.80	75.50	77.90	74.19	74.20
		1.5	66.54	67.36	64.68	63.67	55.00	57.17	52.92	51.08	74.78	77.23	72.82	70.45
		1.8	69.73	69.61	67.39	60.69	58.20	59.80	55.36	47.87	78.27	80.48	75.89	66.73
	U_g	1.2	49.68	53.22	46.03	51.75	37.79	43.17	34.37	40.34	66.70	70.98	64.24	67.36
		1.5	49.10	52.34	43.75	49.02	37.68	43.31	32.50	37.94	65.72	70.48	62.12	63.07
		1.8	52.49	53.55	45.70	46.96	41.39	45.53	35.05	35.65	70.62	74.58	66.03	60.01
Under H_{1B}	E_1	1.2	63.68	65.29	60.93	64.20	78.76	81.12	76.60	79.45	94.25	95.48	93.42	94.84
		1.5	62.59	63.97	58.40	61.04	77.57	79.78	74.02	75.93	93.66	95.03	92.14	92.79
		1.8	64.50	64.54	58.94	58.38	79.66	80.83	74.82	72.72	94.86	95.79	92.84	90.90
	E_2	1.2	91.25	91.70	90.63	91.04	98.64	98.75	98.47	98.50	99.98	99.99	99.98	99.98
		1.5	90.31	90.73	89.18	69.47	98.27	98.45	97.95	97.75	99.96	99.97	99.94	99.93
		1.8	91.09	91.29	89.44	86.91	98.50	98.58	98.06	96.73	99.99	99.99	99.98	99.90
	U_g	1.2	76.34	79.25	72.91	78.61	90.10	92.53	87.61	91.94	99.85	99.90	99.82	99.87
		1.5	75.08	78.18	68.95	76.05	89.23	91.98	84.67	89.96	99.82	99.88	99.71	99.77
		1.8	76.68	78.55	68.12	73.58	90.33	92.20	83.85	87.93	99.87	99.91	99.73	99.64

Table 7. The powers (in %) under Donner's model at the nominal level of $\alpha = 0.05$.

(m, n)		ρ	$g = 2$				$g = 4$				$g = 8$			
			Q_{LR}	Q_W	Q_S	Q_{GS}	Q_{LR}	Q_W	Q_S	Q_{GS}	Q_{LR}	Q_W	Q_S	Q_{GS}
Under H_{1A}	E_1	0.1	39.38	40.53	38.55	39.65	28.46	31.70	27.45	28.93	39.27	44.79	37.68	39.85
		0.5	34.98	36.19	34.41	34.25	25.33	28.74	24.30	24.71	34.47	40.73	32.78	33.37
		0.7	33.13	34.26	32.51	32.48	23.87	27.27	22.89	23.41	32.10	38.64	30.53	31.45
	E_2	0.1	67.81	68.55	67.36	67.41	55.48	57.43	54.69	54.84	75.26	77.48	74.41	74.48
		0.5	60.20	60.90	59.91	59.50	47.65	49.65	47.02	46.73	65.83	68.64	64.98	64.69
		0.7	57.29	57.97	57.00	56.66	44.95	46.93	44.25	44.15	62.33	65.35	61.41	61.41
	U_g	0.1	49.44	52.49	47.99	51.55	37.92	42.85	35.98	40.43	66.66	70.72	65.05	67.43
		0.5	43.89	47.12	42.42	45.12	33.36	38.73	31.38	34.75	58.11	63.13	56.45	58.31
		0.7	41.47	44.73	39.95	42.87	30.82	36.20	28.83	32.48	54.20	59.60	52.47	54.61
Under H_{1B}	E_1	0.1	64.00	64.84	63.19	64.59	78.84	80.71	78.09	79.53	94.30	95.45	93.94	94.69
		0.5	57.48	58.69	56.83	56.76	72.05	74.73	71.05	71.08	90.23	92.21	89.48	89.52
		0.7	54.77	55.82	54.00	53.91	68.75	71.52	67.58	67.69	87.84	90.22	86.97	87.27
	E_2	0.1	91.43	91.72	91.25	91.31	98.53	98.67	98.46	98.50	99.98	99.98	99.98	99.98
		0.5	86.11	86.49	85.91	85.50	96.28	96.56	96.16	95.97	99.85	99.87	99.84	99.83
		0.7	83.63	83.95	83.42	83.08	95.06	95.44	94.90	94.77	99.68	99.73	99.66	99.64
	U_g	0.1	76.36	78.83	75.10	78.32	90.02	92.30	88.97	91.55	99.88	99.91	99.85	99.89
		0.5	69.42	72.58	68.03	70.66	84.68	88.06	83.14	85.82	99.45	99.61	99.37	99.44
		0.7	66.47	69.60	64.85	67.86	81.80	85.58	79.97	83.27	99.04	99.33	98.92	99.07

The first example consists of a subset of 214 children who were admitted because of acute otitis media with effusion (OME), and were randomized into two groups, respectively treated with cefaclor and amoxicillin[40]. Table 8 shows the number of cured ears in 173 children at 42 d.

Based on the goodness-of-fit test procedure by Zhou & Ma[39], it shows that both Rosner's and Donner's models are suitable for this dataset (all p-values for goodness-of-fit test $\gtrsim 0.9$), with Rosner's model yielding slightly lower AIC (274.1305 under Rosner's model vs 274.1406 under Donner's model). Therefore, Rosner's model was selected to perform the equal proportion test. The constrained and unconstrained MLEs, along with the statistics and p-values under Rosner's model, can be found in Table 9. It can be seen that all the p-values from the three likelihood-based tests, and from the GEE test are of similar magnitude ($\gtrsim 0.8$), and much larger than the nominal level of $\alpha = 0.05$. Therefore, we fail to reject the null hypothesis H_0 : $\pi_1 = \pi_2$, indicating no significant difference in the proportion of the cured ears between the two treatments (cefaclor versus amoxicillin). This conclusion aligns with the findings in the original study[40], which reported that '*by 42 d after entry, the percentage of children whose ears were without effusion or "improved" was equal in both treatment groups (68.9% in the cefaclor group and 67.5% in the amoxicillin group)*', despite overlooking the intra-subject correlation. However, by taking the intra-subject correlation into account in the likelihood-based tests and the GEE test, stronger evidence for the equal proportion of cured ears across treatment groups is observed, as reflected by very large p-values.

The second example involves the ophthalmologic study conducted in the Massachusetts Eye and Ear Infirmary between 1970 and 1979, where data were collected from an outpatient population of patients with retinitis pigmentosa (RP), and their normal relatives[41]. Patients were classified into four types of genetic groups: (1) autosomal dominant RP (DOM); (2) autosomal recessive RP (AR); (3) sex-linked RP (SL); and (4) isolate RP (ISO). A subset of 218 patients aged 20–39 was originally analyzed by Rosner using the constant R model[1], where an eye was considered affected if the best corrected Snellen visual acuity (VA) was 20/50 or worse, and normal if VA was 20/40 or better. Table 10 presents the distribution of the number of affected eyes for 216 patients in the four genetic groups who had complete information for VA on both eyes. This example can be considered as a special case in the combined data framework where unilateral observations are absent.

Table 8. Number of cured ears at 42 d in children treated with cefaclor and amoxicillin.

No. of cured ears	Treatment		Total
	Cefaclor	Amoxicillin	
0	9	7	16
1	7	5	12
2	23	13	36
Total	39	25	64
0	20	19	39
1	34	36	70
Total	54	55	109

Using the goodness-of-fit test results presented in Zhou & Ma[39], it can be seen that Donner's model is much superior to Rosner's model as reflected by one magnitude larger p-values, and smaller AIC (443.7967 under Donner's model vs 449.9490 under Rosner's model). Therefore, Donner's model was selected to perform the equal proportion test. The constrained and unconstrained MLEs, along with the statistics and p-values under Donner's model, are presented in Table 11. As can be seen, all the p-values are smaller than 0.05, with the p-values from the score and GEE tests being close to each other. Therefore, the null hypothesis H_0: $\pi_1 = \pi_2 = \pi_3 = \pi_4$ is rejected, and it is concluded that there is an overall difference in proportions of the affected eyes across the four genetic groups. This finding is consistent with the original paper by Rosner[1], which also noted that the overall difference was completely attributed to the differences between the SL group, and each of the other three groups.

Discussion and conclusions

In the present study, the GEE approach is considered as an alternative to three likelihood-based methods (likelihood ratio, Wald-type, and score statistics) for testing homogeneity of proportions across g ≥ 2 groups for the combined unilateral and bilateral data.

Table 9. Constrained and unconstrained MLEs along with statistics and *p*-values under Rosner's model for the OME dataset.

	Constrained MLEs		Unconstrained MLEs	
	Cefaclor	Amoxicillin	Cefaclor	Amoxicillin
$\hat{\pi}_i$	$\hat{\pi}_0 = 0.6482$		$\hat{\pi}_1 = 0.6528$	$\hat{\pi}_2 = 0.6425$
$\hat{R}$	$\hat{R}_0 = 1.3182$		$\hat{R} = 1.3172$	
$\hat{\rho}$	$\hat{\rho}_0 = 0.5862$		$\hat{\rho}_1 = 0.5964$	$\hat{\rho}_2 = 0.5699$
	Q_{LR}	Q_W	Q_S	Q_{GS}
Statistic	0.0394	0.0391	0.0395	0.0265
p-value	0.8426	0.8432	0.8424	0.8706

Table 10. Number of affected eyes for patients in four genetic groups.

No. of affected eyes	Genetic type				Total
	DOM	AR	SL	ISO	
0	15	7	3	67	92
1	6	5	2	24	37
2	7	9	14	57	87
Total	28	21	19	148	216

Table 11. Constrained and unconstrained MLEs along with statistics and *p*-values under Donner's model for the retinitis pigmentosa dataset.

		DOM	AR	SL	ISO
Constrained MLEs	$\hat{\pi}_i$	$\hat{\pi}_0 = 0.4884$			
	$\hat{\rho}$	$\hat{\rho}_0 = 0.6572$			
Unconstrained MLEs	$\hat{\pi}_i$	$\hat{\pi}_1 = 0.3625$	$\hat{\pi}_2 = 0.5455$	$\hat{\pi}_3 = 0.7926$	$\hat{\pi}_4 = 0.4658$
	$\hat{\rho}$	$\hat{\rho} = 0.6416$			
		Q_{LR}	Q_W	Q_S	Q_{GS}
Statistic		12.0385	16.3267	11.3158	10.6890
p-value		0.0073	0.0010	0.0101	0.0135

The likelihood-based methods are revisited under two statistical models: i) Rosner's constant R model; and ii) Donner's equal correlation ρ model, accounting for intra-subject correlations in the bilateral portion, and introduce the GEE method along with its implementation in SAS GENMOD procedure for the combined organ data.

Monte Carlo simulations evaluate the empirical type I error rates, and powers of each method under various sample sizes and parameter configurations. Results show that the GEE and score tests exhibit comparable performance, outperforming the likelihood ratio and Wald-type tests in controlling type I error. All tests demonstrate similar power, though the likelihood ratio, and Wald-type tests are slightly inflated. These findings confirm that the GEE test performs at least as well as the score test and revalidate previous results reported by Ma and others[7,19]. Importantly, the GEE method offers additional flexibility, allowing incorporation of continuous covariates and more complex model structures, which provides an advantage over the score test.

Applications to two real datasets from otolaryngologic and ophthalmologic studies illustrate these methods. Goodness-of-fit tests guide model selection for the likelihood-based methods, showing that Rosner's model is preferred for the OME dataset, while Donner's model better fits the retinitis pigmentosa dataset. Inference from both the GEE method and the likelihood-based methods after model selection are consistent with the original study results.

In conclusion, the GEE and score tests are recommended for homogeneity testing for the combined unilateral and bilateral data. Both methods provide robust type I error control, and strong power across different scenarios. For extended analyses involving continuous covariates or multiple explanatory variables, the GEE approach is preferred. The score test is computationally efficient, particularly for large samples or numerous groups, and is well suited for exact tests, such as Fisher's exact or permutation tests, when data are sparse, which is less straightforward with standard software implementation of GEE.

Author contributions

The authors confirm their contributions to the paper as follows: study conception and design: Ma CX; analysis and interpretation of results: Zhou J, Ma CX; draft manuscript preparation: Zhou J. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The data presented in this study are openly available in the studies by Rosner and Mandel et al.[1,40]. The full source code used for the simulation studies and real data analyses, including scripts for generating and running SAS procedures, is publicly available at: https://github.com/pmdnticc/Homogeneity_Test_Combined_Data. Detailed instructions for reproducing the analyses are provided in the accompanying README files.

Acknowledgments

Not applicable.

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 16 August 2025; Revised 19 December 2025; Accepted 13 January 2026; Published online 28 March 2026

References

[1] Rosner B. 1982. Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes. *Biometrics* 38:105–114

[2] Dallal GE. 1988. Paired bernoulli trials. *Biometrics* 44:253–257

[3] Donner A. 1989Statistical methods in ophthalmology: an adjusted chi-square approach. *Biometrics* 45:605–611

[4] Thompson JR. 1993. The chi 2 test for data collected on eyes. *British Journal of Ophthalmology* 77:115–117

[5] Ma C, Shan G, Liu S. 2015. Homogeneity test for correlated binary data. *PLoS One* 10:e0124337

[6] Qiu SF, Poon WY, Tang ML, Tao JR. 2019. Construction of confidence intervals for the risk differences in stratified design with correlated bilateral data. *Journal of Biopharmaceutical Statistics* 29:446–467

[7] Ma CX, Wang K. 2021. Testing the homogeneity of proportions for combined unilateral and bilateral data. *Journal of Biopharmaceutical Statistics* 31:686–704

[8] Qiu SF, Tao JR. 2022. Confidence intervals for assessing equivalence of two treatments with combined unilateral and bilateral data. *Journal of Applied Statistics* 49:3414–3435

[9] Qiu SF, Liu QS, Ge Y. 2023. Confidence intervals of proportion differences for stratified combined unilateral and bilateral data. *Communications in Statistics-Simulation and Computation* 52:3839–3862

[10] Li Y, Li Z, Mou K. 2023. Homogeneity test of many-to-one relative risk ratios in unilateral and bilateral data with multiple groups. *Axioms* 12:333

[11] Liu Y, Li Z, Mou K, Du J. 2025. Testing the equality of response rate functions for paired binary data with multiple groups. *Statistical Methods in Medical Research* 34:131–149
[12] M'lan CE, Chen MH. 2015. Objective Bayesian inference for bilateral data. *Bayesian Analysis* 10:139–170
[13] Zhao H, Wang X, Bian J, Chen S, Li Z. 2023. Homogeneity test of response rate functions in bilateral correlated data under dallal's model. *Complexity* 2023:1–22
[14] Sun S, Li Z, Ai M, Jiang H. 2022. Risk difference tests for stratified binary data under Dallal's model. *Statistical Methods in Medical Research* 31:1135–1156
[15] Sun S, Li Z, Jiang H. 2024. Homogeneity test and sample size of risk difference for stratified unilateral and bilateral data. *Communications in Statistics-Simulation and Computation* 53:4209–4232
[16] Sun S, Li Z, Mou K. 2025. Interval estimation of common risk difference for stratified unilateral and bilateral data. *Journal of Biopharmaceutical Statistics* 35:85–105
[17] Ma CX, Liu S. 2017. Testing equality of proportions for correlated binary data in ophthalmologic studies. *Journal of Biopharmaceutical Statistics* 27:611–619
[18] Mou K, Li Z. 2021. Homogeneity test of many-to-one risk differences for correlated binary data under optimal algorithms. *Complexity* 2021:6685951
[19] Ma CX, Wang H. 2023. Testing the equality of proportions for combined unilateral and bilateral data under equal intraclass correlation model. *Statistics in Biopharmaceutical Research* 15:608–617
[20] Cheng J, Li Z, Mou K. 2025. Statistical analysis of a generalized linear model for bilateral correlated data under Donner's model. *Axioms* 14:500
[21] Tang NS, Tang ML, Qiu SF. 2008. Testing the equality of proportions for correlated otolaryngologic data. *Computational Statistics and Data Analysis* 52:3719–3729
[22] Ying GS, Maguire MG, Glynn R, Rosner B. 2017. Tutorial on biostatistics: linear regression analysis of continuous correlated eye data. *Ophthalmic Epidemiology* 24:130–140
[23] Ying GS, Maguire MG, Glynn R, Rosner B. 2018. Tutorial on biostatistics: statistical analysis for correlated binary eye data. *Ophthalmic Epidemiology* 25:1–12
[24] Schall R. 1991. Estimation in generalized linear models with random effects. *Biometrika* 78:719–727
[25] Breslow NE, Clayton DG. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88:9–25
[26] McCulloch CE, Searle SR. 2000. *Generalized, Linear, and Mixed Models*. Hoboken, NJ: John Wiley & Sons. 335 pp. doi: 10.1002/0471722073
[27] Stroup WW, Ptukhina M, Garai J. 2024. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. 2nd Edition. New York: Chapman and Hall/CRC. 668 pp. doi: 10.1201/9780429092060
[28] Liang KY, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22
[29] Zeger SL, Liang KY. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 121–130
[30] Zeger SL, Liang KY, Albert PS. 1988. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44:1049–1060
[31] Liang KY, Zeger SL. 1993. Regression analysis for correlated data. *Annual Review of Public Health* 14:43–68
[32] Diggle PJ, Heagerty P, Liang KY, Zeger SL. 2002. *Analysis of Longitudinal Data*. Oxford, UK: Oxford University Press. 348 pp. doi: 10.1093/oso/9780198524847.001.0001
[33] Wang YG, Carey V. 2003. Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. *Biometrika* 90:29–41
[34] Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, et al. 2010. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* 21:467–474
[35] Fitzmaurice GM, Laird NM, Ware JH. 2011. *Applied Longitudinal Analysis*. Hoboken, NJ: John Wiley & Sons. 711 pp. doi: 10.1002/9781119513469
[36] Zhang X, Ma CX. 2025. Analysis of bilateral and unilateral data: a comparative review of model-based and MLE-based methods for the homogeneity test of proportions. *Journal of Data Science* 23:638–647
[37] SAS Institute Inc. 2017. *SAS/STAT 14.3 User's Guide: The GENMOD Procedure*, volume 46. Cary, NC: SAS Institute Inc. 213 pp. https://support.sas.com/documentation/onlinedoc/stat/143/genmod.pdf
[38] Boos DD. 1992. On generalized score tests. *The American Statistician* 46:327–333
[39] Zhou J, Ma CX. 2025. Goodness-of-fit tests for combined unilateral and bilateral data. *Mathematics* 13:2501
[40] Mandel EM, Bluestone CD, Rockette HE, Blatter MM, Reisinger KS, et al. 1982. Duration of effusion after antibiotic treatment for acute otitis media: comparison of cefaclor and amoxicillin. *The Pediatric Infectious Disease Journal* 1:310–316
[41] Berson EL, Rosner B, Simonoff E. 1980. Risk factors for genetic typing and detection in retinitis pigmentosa. *American Journal of Ophthalmology* 89:763–775