

# Test for Homogeneity of Odds Ratios Using $U$ -Statistics

Qi Wei<sup>1</sup>, Dejian Lai<sup>2\*</sup>

<sup>1</sup>Surgery Core Research, Baylor College of Medicine, Houston, USA

<sup>2</sup>Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, USA

Email: \*Dejian.lai@uth.tmd.edu

**How to cite this paper:** Wei, Q. and Lai, D.J. (2019) Test for Homogeneity of Odds Ratios Using  $U$ -Statistics. *Open Journal of Statistics*, 9, 347-360.

<https://doi.org/10.4236/ojs.2019.93024>

**Received:** May 9, 2019

**Accepted:** June 15, 2019

**Published:** June 18, 2019

Copyright © 2019 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

---

## Abstract

There are a few statistics testing the homogeneity of odds ratios across strata. Asymptotic statistics lose their power in the “sparse-data” setting. Both asymptotic statistics and exact tests have low power when the sample sizes are small. We created a set of  $U$  statistics and compared them with some existing statistics in testing homogeneity of  $OR$  at different data settings. We evaluated their performance in terms of the empirical size and power via Monte Carlo simulations. Our results showed that two of the  $U$ -statistics under our study had higher power for testing homogeneity of odds ratios for 2 by 2 contingency tables. The application of the tests was illustrated in two real examples.

## Keywords

Homogeneity Test, Odds Ratio,  $U$ -Statistics

---

## 1. Introduction

Odds ratio is commonly used in the analysis of association of two factors that both have two categories. In epidemiological studies and clinical trials, these two factors usually refer to the exposure (treatment/intervention/risk) factor  $X$  and the outcome factor  $Y$  respectively. The association between  $X$  and  $Y$ , however, could be modified or confounded by a third factor  $Z$ . For example, in a multi-center clinical trial, factor  $Z$  could be the center. Each center is corresponding to a stratum of  $Z$ . Because the presence of the heterogeneity of odds ratios may lead to different methods of analysis, researchers usually want to test whether the odds ratios are homogeneous across the strata of the factor  $Z$  or not. This type of tests is called the tests for homogeneity of the odds ratios, or the tests for null interaction.

A few procedures have been developed for testing the homogeneity of odds ratios. They are usually categorized into two classes: exact tests and asymptotic tests. Most of the asymptotic statistics were derived for “large-stratum” settings, where the sample size is large, and the number of strata is small. Liang and Self [1] developed two asymptotic statistics—score statistics for the “sparse-data” setting, where there are many cells with small counts and/or zeros. The asymptotic tests could be poor if there were some cells with small cell counts in the contingency tables even when the sample size was quite large. The empirical sizes of the asymptotic tests were shown to be conservative when the data were “small-stratum” setting [2]. Jones *et al.* [3] showed in their simulation study that, in the “sparse-data” setting, five asymptotic tests suffered in both empirical size and power. Zelen [4] constructed an exact test for homogeneity that employs the ordering principle for a single 2 by 2 table, which was reexamined by Halperin *et al.* [5]. Algorithms were developed for the Zelen exact test later [6] [7] [8] [9]. Hirji *et al.* [10] showed that their algorithm for the Zelen statistic was more versatile and efficient in terms of the accuracy and the usage of memory in the computer computation. Hirji *et al.* [10] also constructed five other exact statistics from their asymptotic counterparts, which are score, likelihood ratio, Pearson  $\chi^2$  and mixture model  $\chi^2$  tests. The exact tests are computationally intensive, particularly when the cell counts are large. Reis, Hirji and Afifi [11] examined the performance of the empirical power of six asymptotic statistics and six exact tests mentioned above; the results of the study showed that the power of both exact and asymptotic tests was low when sample sizes were small even with relatively large heterogeneity among odds ratios. Baghri *et al.* [12] compared three tests of homogeneity of odds ratios. A recent study investigated profile likelihood tests for common odds ratios in meta-analysis [13].

Our study presented in this paper compared a class of  $U$ -statistics with the established tests such as the Zelen test and the Breslow-Day test.

## 2. $U$ -Statistics

The statistics of the form

$$U = [l!(n-l)!/n!] \sum_{(n,l)} h(X_{i_1}, \dots, X_{i_l}) \quad (1)$$

are known as  $U$ -statistics, where  $\{X_{i_1}, \dots, X_{i_l}\}$  is a set of  $l$ -subset of  $\{1, \dots, n\}$ ; the sum  $\sum_{(n,l)}$  is taken over all subsets  $1 \leq i_1 < \dots < i_l \leq n$  of  $\{1, 2, \dots, n\}$ ;  $h$  is the

kernel function and symmetric in its arguments. In our study, we investigated the applicability of this class of statistics with  $l = 2$  in testing homogeneity of odds ratios among 2 by 2 tables.  $U$ -statistics were first identified as a minimum-variance unbiased estimator by Halmos [14] and were named by Hoeffding [15], who demonstrated the asymptotic normality of this class of statistics in his seminal paper [15]. Many statistics in common use are members of this class, including the sample mean, the sample variance and the sample covariance [16] [17].

**Table 1** shows a 2 by 2 table of a stratum  $k$ .

Assume that  $a_k$  and  $b_k$  are counts of independent binomial outcomes from number  $n_k$  and  $m_k$  of trials with or without exposure at the stratum  $k$  respectively;  $N_k$  is the total sample size of the stratum  $k$ ; and the tables are independent among strata. The commonly used estimate of the odds ratio of the  $k$ th stratum is expressed as:  $\hat{\psi}_k = (a_k / (n_k - a_k)) / (b_k / (m_k - b_k))$ . We want to test whether the odds ratios are homogeneous among all  $K$  strata (or all levels of the third variable  $Z$ ), that is to test  $H_0: \psi_1 = \psi_2 = \dots = \psi_K = \psi$  against  $H_a: \psi_i \neq \psi_j$  for at least one pair of  $(i, j)$ , where  $i, j = 1, \dots, K, i \neq j$ .

Our study evaluated a class of weighted  $U$ -statistics  $\sum_{i < j} w_{ij} h(\hat{\psi}_i, \hat{\psi}_j)$  as well as a class of unweighted  $U$ -statistics  $(2!(K-2)!/K!) \sum_{i < j} h(\hat{\psi}_i, \hat{\psi}_j)$ , where  $\hat{\psi}_i$  and  $\hat{\psi}_j$  are the estimates of the odds ratios in the  $i$ th and  $j$ th stratum of  $Z$ ,  $h(\hat{\psi}_i, \hat{\psi}_j)$  is a function of  $\hat{\psi}_i$  and  $\hat{\psi}_j$ , and  $w_{ij}$  is the weight associated with  $h(\hat{\psi}_i, \hat{\psi}_j)$ . Based on the simulation results, we only focus our attention on the following two statistics in this paper:

$$U3: (2!(K-2)!/K!) \sum_{i < j} |\log \hat{\psi}_i - \log \hat{\psi}_j|, \tag{2}$$

$$WU3: \sum_{i < j}^K w_{ij} |\log \hat{\psi}_i - \log \hat{\psi}_j|. \tag{3}$$

The base of log in all formula is  $e$ . The sample distribution of the estimated odds ratio is highly skewed when the sample size is small or moderate. Because of this, we used the natural logarithm of  $\hat{\psi}$  in  $U3$  and  $WU3$  to reduce the skewness. Consider that a large stratum offers more accurate estimate for the odds ratio, a weight was selected for each  $h(\hat{\psi}_i, \hat{\psi}_j)$  in expression (3), which is proportional to the stratum's size. It has the following form:

$$w_{ij} = (w_i w_j) / \sum w_k \tag{4}$$

where  $w_i = 1/\text{var}(\log(\hat{\psi}_i))$ . The log transform of the sample odds ratio has an asymptotic variance in a simple form, which is,

$\text{var}(\log \hat{\psi}_k) \approx 1/a_k + 1/(n_k - a_k) + 1/b_k + 1/(m_k - b_k)$ . If there was any cell count that equals to zero, the odds ratio estimate  $\hat{\psi}_k$  and  $\text{var}(\log \hat{\psi}_k)$  from the above formula would be undefined. We added 0.5 to each cell count of that table in the calculations to get the amended estimators [18]. In our simulation, we studied other forms of  $U$ -statistics. Due to the relatively poor performance of the others, we only report the results of the  $U3$  and the  $WU3$  and their comparisons to the Zelen's test and the Breslow-Day test in this paper. Detailed results are available upon request.

**Table 1.**  $2 \times 2$  table of  $X$  and  $Y$  at  $Z = k$ .

	$Y = 1$	$Y = 0$	
$X = 1$	$a_k$	$n_k - a_k$	$n_k$
$X = 0$	$b_k$	$m_k - b_k$	$m_k$
	$t_k$	$N_k - t_k$	$N_k$

### 3. Simulation Study

A total of 10,000 data sets were simulated using the SAS subroutine RANBIN. Each data set contains pre-specified  $K$  sets of 2 by 2 tables. The cell counts  $a_k$  and  $b_k$  were independently generated from binomial distributions  $(n_k, p_{1k})$  and  $(m_k, p_{0k})$ , where  $p_{xk} = P(Y = 1 | X = x, Z = k)$  is the probability of  $Y = 1$  when  $X = x$  ( $x = 1, 0$ ) in the  $k$ th stratum. Each set of the tables was simulated with a given  $n_k, m_k$ , the number of the strata  $K$  and the odds ratios. For a given odd ratio  $\psi_k$  and a binomial proportion  $p_{0k}, p_{1k}$  was calculated by solving:

$$\psi_k = \left[ \frac{p_{1k}}{1 - p_{1k}} \right] / \left[ \frac{p_{0k}}{1 - p_{0k}} \right].$$

Following the previous simulation study by Reis, Hirji and Afifi [11], five factors that might influence the sizes and power of the test statistics were studied. These five factors are: 1) the sample size for  $X = 1$  ( $n_k$ ); 2) the ratio of the sample sizes of  $X = 0$  and  $X = 1$  ( $m_k:n_k$ ); 3) the probability of  $Y = 1$  among  $X = 0$  ( $p_{0k}$ ); 4) the odds ratio (OR)  $\psi_k$ ; 5) the number of  $2 \times 2$  tables,  $K$  (strata). In order to study the performance of the statistics, different values of the above five factors (parameters) were chosen to simulate different data sets. In choosing the values of  $n_k, m_k, n_k, p_{0k}, \psi_k$ , and  $K$ , we took the following factors into account: the situations in practice, the characteristics of  $U$ -statistics and the design of the simulation study by Reis, Hirji and Afifi [11]. When the effect of one factor was studied, the other four factors' values were fixed.

We compared the performance of the  $U$ -statistics with the Breslow-Day statistic and Zelen's exact test in our simulation study. A C++ program was written to calculate these statistics' exact  $P$ -values, empirical sizes and power. The empirical size was calculated as the percentage of times that the test rejected the null hypothesis of a common odds ratio at a pre-specified  $\alpha$  level among 10,000 tests that were simulated with same odds ratios among  $K$  tables. The empirical power was calculated as the percentage of times that a test rejected the null hypothesis of a common odds ratio at a prescribed  $\alpha$  level when data were simulated under alternative hypotheses. Because the  $U$ -statistics studied here are functions of the sums across all the absolute distances between all possible pairs of the estimated odds ratios in log scale, a large value of  $U$  statistics indicates the heterogeneity of the odds ratio.

Theoretically, under suitable conditions,  $\frac{[T - E(T)]}{\sqrt{\text{var}(T)}}$  will be asymptotically following  $N(0, 1)$  as  $K \rightarrow \infty$ , where  $T$  represents a  $U$ -statistic. In our study, the sample mean and the sample variance of 10,000 statistics under the null hypothesis were used to estimate the  $E(T)$  and the  $\text{var}(T)$ . In an actual application, the sample mean and variance may be estimated as suggested in our application section. In the simulation, one would use the result from expression [9] for the variance. In so doing, one would get a different numerical value of the variance for each realization, which was inconsistent with our assumption of common variance. The effect of unequal variance is still under investigation.

## 4. Results of the Simulation Study

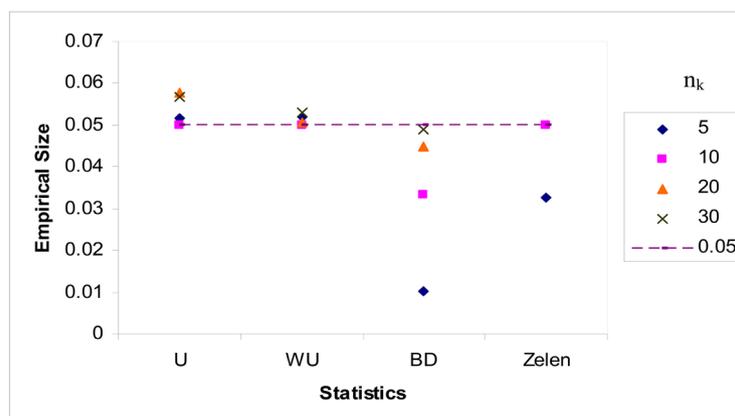
### 4.1. Empirical Sizes of the Tests

The five factors affected the test statistics differently. The empirical size of the Breslow-Day test was improved (moved closer to the predefined  $\alpha$  level) as the values of  $n_k$ ,  $m_k:n_k$ ,  $p_{0k}$  and odds ratios increased but diverged from the pre-specified  $\alpha$  level when the number of stratum increased. A weak trend was observed that the empirical size of  $U3$  and  $WU3$  moved closer and then diverged from the pre-specified  $\alpha$  level when the sample size  $n_k$  increased (Figure 1). Similar results were observed in studying the effect of ratio  $m_k:n_k$ ; their empirical sizes were improved when  $m_k:n_k$  increased to moderate value ( $m_k:n_k$  equal to 2 or 3), they then diverged from the predefined significant level 0.05 if the value of  $m_k:n_k$  was higher (Figure 2). The empirical sizes of  $U$ -statistics increased as the value of the odds ratio increased. The probability  $p_{0k}$  had the least impact on the empirical sizes of the two  $U$ -statistics.

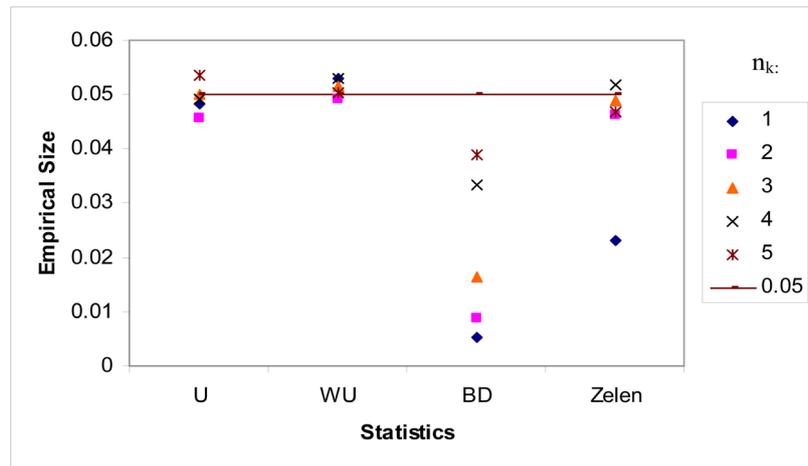
The number of strata  $K$  had an apparent effect on the sizes of  $U$ -statistics; their empirical sizes were improved as  $K$  increased (Figure 3), even when the total sample size remained the same (Figure 4). The empirical size of Breslow-Day test diverged from the predefined nominal size of 5% as the value of  $K$  increased. The empirical size of Zelen test improved as  $K$  and total sample size increased but diverged from nominal size of 5% when the number of strata increased without increasing the total sample size.

### 4.2. Empirical Power of the Statistics

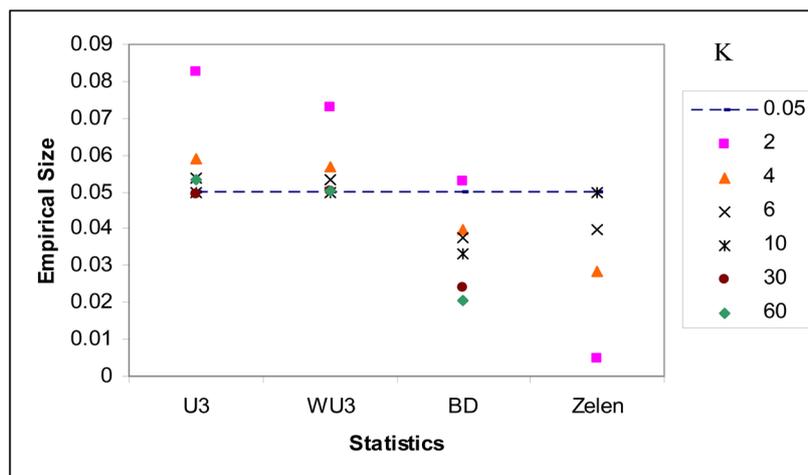
Seven settings of heterogeneous odds ratios were evaluated as alternative hypotheses in our study. However, in this article, we only reported the empirical powers from the scenario that the alternative odds ratios were generated following the pattern of 1, 2, 3, 7. That is, 25% of the generated tables under  $H_a$  have odds ratios being 1, 2, 3 and 7, respectively. In order to show the effects of different factors on the test statistics, we also simulated the critical values based on these factors (Figures 5-11).



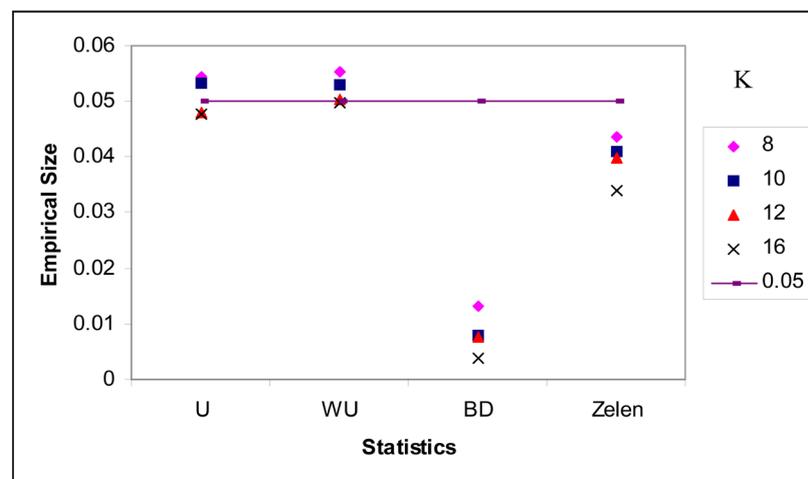
**Figure 1.** The effect of  $n_k$  on empirical size, Sparse-strata setting, between strata balanced: ( $p_{0k} = 0.1$ ,  $K = 10$ ,  $m_k:n_k = 1$ ,  $OR = 2$ ).



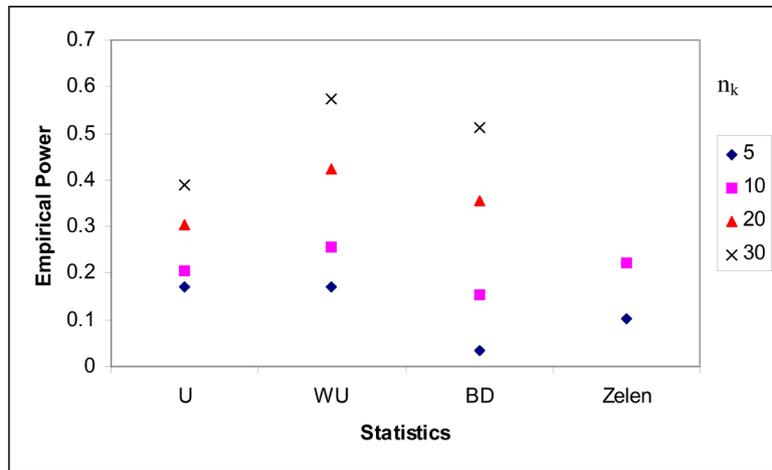
**Figure 2.** The effect of ratio  $m_k:n_k$  on empirical size, between strata balanced setting: ( $p_{0k} = 0.05$ ,  $K = 10$ ,  $n_k = 10$ ,  $OR = 2$ ).



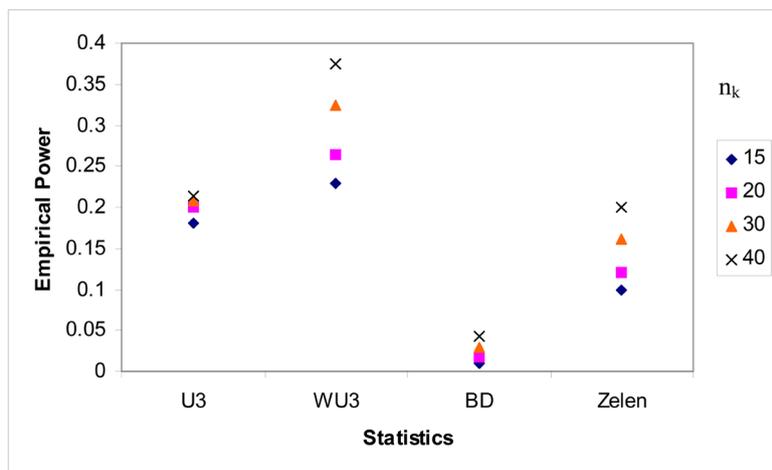
**Figure 3.** The effect of the number of strata  $K$  on empirical size, between strata balanced and within strata balanced: ( $p_{0k} = 0.05$ ,  $n_k = 10$ ,  $m_k:n_k = 1$ ,  $OR = 2$ ).



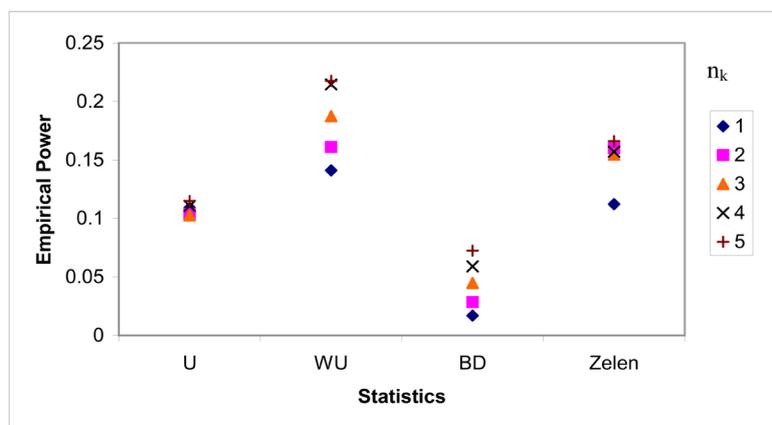
**Figure 4.** The effect of the number of strata  $K$  on empirical size, between strata unbalanced and within strata balanced: ( $p_{0k} = 0.1$ ,  $n_k = 10$ ,  $m_k:n_k = 1$ ,  $OR = 2$ ).



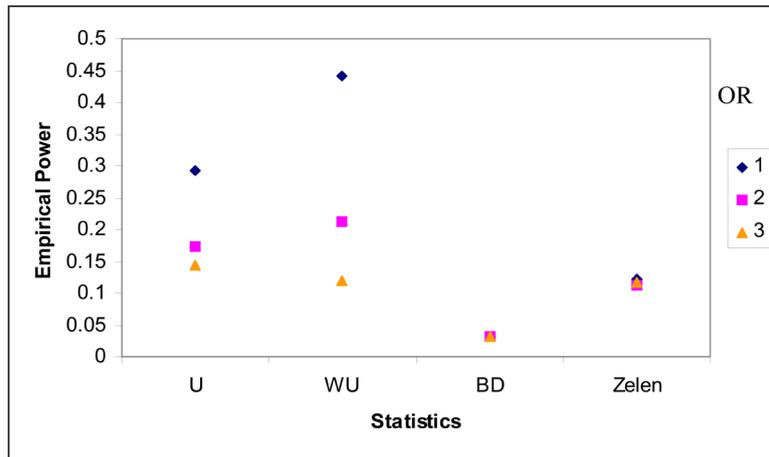
**Figure 5.** The effect of  $n_k$  on empirical power, between strata balanced: ( $p_{0k} = 0.1, K = 10, m_k \cdot n_k = 1, Ha7$  shown in **Table 1**).



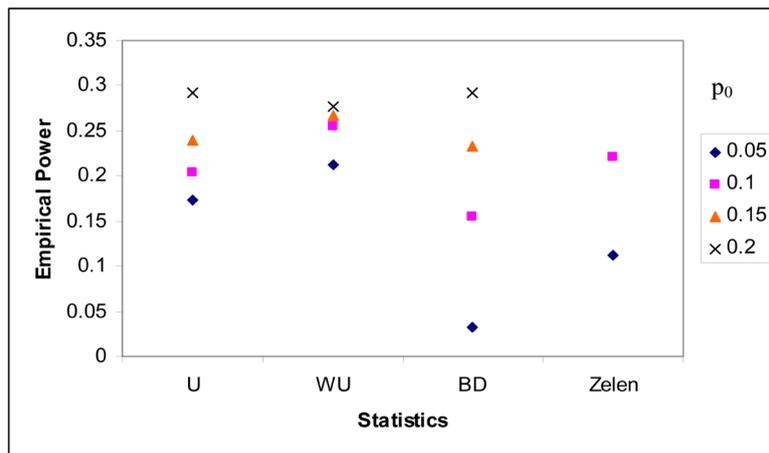
**Figure 6.** The effect of  $n_k$  on empirical power, between strata unbalanced and within strata balanced setting: Two large strata with  $OR = 2$ , and 7,  $n_k$  changed as shown in figure, Eight small strata with  $OR = 1, 2, 3, 7, 1, 2, 3, 7, n_k = 5$ . ( $p_{0k} = 0.05, K = 10, m_k \cdot n_k = 1, Ha7$  shown in **Table 1**).



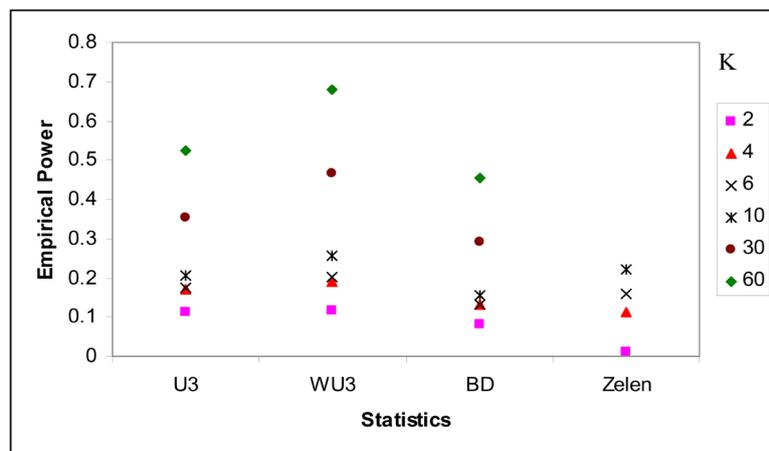
**Figure 7.** The effect of ratio  $m_k/n_k$  on empirical power, between strata balanced setting: ( $p_{0k} = 0.05, K = 10, n_k = 10, OR = 2, Ha7$  shown in **Table 1**).



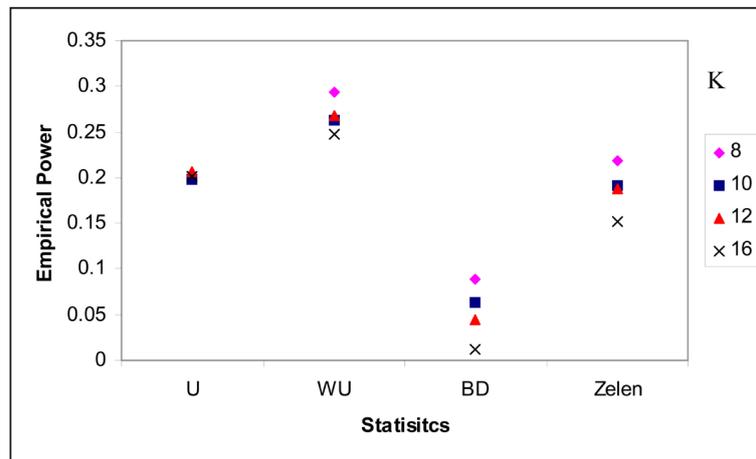
**Figure 8.** The effect of the value of odds ratio on empirical power, between strata balanced and within strata balanced, ( $p_{0k} = 0.05$ ,  $n_k = 10$ ,  $m_k:n_k = 1$ ,  $K = 10$ , Ha7 shown in Table 1).



**Figure 9.** The effect of the value of  $p_{0k}$  on empirical power, between strata balanced and within strata balanced, ( $n_k = 10$ ,  $m_k:n_k = 1$ ,  $K = 10$ ,  $OR = 2$ , Ha7 shown in Table 1).



**Figure 10.** The effect of the number of strata  $K$  on empirical power, between strata balanced and within strata balanced: ( $p_{0k} = 0.05$ ;  $n_k = 10$ ;  $m_k:n_k = 1$ ;  $OR = 2$ , Ha7 shown in Table 1).



**Figure 11.** The effect of the number of strata  $K$  on empirical power, between strata unbalanced and within strata balanced: Total sample size keeps unchanged while changing value of  $K$  ( $p_{0k} = 0.05$ ;  $m_k:n_k = 1$ ;  $OR_i = OR_j = 2$ . 4 small strata with  $n_k = 5$ , and  $(K-4)$  large strata with  $n_k = 240/(2 \times (K - 4))$ , Ha7 shown in **Table 1**).

All the statistics' empirical power increased as  $n_k$  increased (**Figure 5**, **Figure 6**). The value of the ratio  $m_k:n_k$  had very little effect on the empirical power of  $U\mathcal{B}$ ; the empirical power of the weighted  $U$ -statistic and Breslow-Day test were improved as the value of  $m_k:n_k$  increased (**Figure 7**). Increasing the value of the odds ratio under the null hypothesis decreased the empirical power of all statistics studied (**Figure 8**). The empirical power increased as the value of  $p_{0k}$  increased (**Figure 9**). When the number of strata  $K$  increased, together with total sample size increased, the power of the  $U$ -statistics and Zelen's exact test were improved (**Figure 10**); If the total sample size remained the same when  $K$  increased, the empirical sizes of the  $U$ -statistics were improved; the empirical size of the Breslow-Day and Zelen's statistics diverged from 5%; the empirical power of  $U\mathcal{B}$  almost remained unchanged, the others' power decreased (**Figure 11**).

### 5. Application to Published Data

Generally,  $U\mathcal{B}$  and  $WU\mathcal{B}$  performed well in terms of both size and power. Their empirical sizes were stable under various situations and had relatively high power.

With the assumption that the odds ratios of  $K \times 2$  tables are the same; and  $\log(\hat{\psi}_i)$  is normally distributed with variance equal to  $\sigma^2$ , we can derive the estimated expected value of  $U\mathcal{B}$  and  $WU\mathcal{B}$ . They are:

$$\begin{aligned}
 E(U\mathcal{B}) &= (2!(K-2)!/K!) E\left(\sum_{i<j} |\log \hat{\psi}_i - \log \hat{\psi}_j|\right) \\
 &= 2\left(1/(2\sigma\sqrt{\pi})\right) \int_0^\infty x \exp(-x^2/(4\sigma^2)) dx = 2\sigma\sqrt{\pi}
 \end{aligned}
 \tag{5}$$

$$\begin{aligned}
 E(WU\mathcal{B}) &= E\left(\sum_{i<j} w_{ij} |\log \hat{\psi}_i - \log \hat{\psi}_j|\right) \\
 &= E\left(|\log \hat{\psi}_i - \log \hat{\psi}_j|\right) \sum_{i<j} w_{ij} = \sum_{i<j} w_{ij} 2\sigma\sqrt{\pi}
 \end{aligned}
 \tag{6}$$

The variance of  $U3$  can be also expressed as:

$$\text{var}(U3) = [2!(K-2)!/K!][2(K-2)1.436\sigma^2 + (K-2)2\sigma^2]. \quad (7)$$

The variance of  $WU3$  can be also expressed as:

$$\text{var}(WU3) = \left[ 1.436\sigma^2 \sum_{j \neq k} w_{ij}w_{ik} + 2\sigma^2 \sum_{i < j} w_{ij}^2 \right]. \quad (8)$$

To estimate the  $\sigma^2$ , consider the Mantel-Haenzel estimator of common odds ratio  $\hat{\psi}_{MH}$  as a weighted average of odds ratios. Given the weight, we can solve the  $\sigma^2$  as a function of  $\text{var}(\log(\hat{\psi}_{MH}))$ , which is:

$$\sigma^2 = \text{var}(\log(\hat{\psi}_{MH})) \left[ \sum (b_k c_k / N_k) \right]^2 / \sum (b_k c_k / N_k)^2 \quad (9)$$

And the value of  $\text{var}(\log(\hat{\psi}_{MH}))$  can be calculated by the following formula:

$$\begin{aligned} \text{var}(\log(\hat{\psi}_{MH})) = & (\sum G_i P_i) / \left[ 2 \left[ \sum G_i \right]^2 \right] + \sum (G_i Q_i + H_i P_i) / \left[ 2 \sum G_i \sum H_i \right] \\ & + \sum H_i Q_i / 2 \left[ \sum H_i \right]^2 \end{aligned}$$

where  $G_i = a_i d_i / N_i$ ,  $H_i = b_i c_i / N_i$ ,  $P_i = (a_i + d_i) / N_i$ ,  $Q_i = (b_i + c_i) / N_i$ .

To illustrate the application of these two  $U$ -statistics, we applied and compared them to the Breslow-Day statistic and the Zelen statistic in two published data sets: 1) Alcohol assumption data (**Table 2**) are from Statistical Methods in Cancer Research [19], volume 1 page 137, which is an example for introducing the use of Breslow-Day statistic; 2) new drug data (**Table 4**) are from a study of comparing a new drug with a controlled drug among 22 hospital sites [20]. Because of the sparseness of the data, asymptotic tests (Breslow-Day statistic) might not be able to yield an accurate  $p$ -value. The test results for the homogeneity of odds ratios are presented in **Table 2** and **Table 4**. For data set 1), according to the  $p$ -values of the test statistics,  $U3$  rejected the null hypothesis; the  $WU3$ , Breslow-Day, and Zelen statistics accepted the null hypothesis at 0.05 level. For data set 2), Zelen and  $WU3$  rejected the null hypothesis, yet Breslow-Day and  $U3$  accepted the null hypothesis of no difference among odds ratios at 0.05 level (**Tables 2-5**).

## 6. Discussion and Conclusions

To summarize the simulation study that we conducted, the following are some remarks: When the number of strata is not very small, ( $K \geq 6$ ), the empirical size of  $U3$  and  $WU3$  were very stable under various situations and stay very close to the nominal of 0.05. In terms of size and power,  $U3$  and  $WU3$  performed better than the Breslow-Day statistic and the Zelen's exact test. Therefore,  $U3$  and  $WU3$  are considered as better statistics for testing the homogeneity of odds ratios in this situation. The test statistic  $U3$  is recommended when the sample size is the same in each stratum, the number of strata is large and the sample size in each stratum is not large. Otherwise,  $WU3$  is recommended.

Breslow-Day test is conservative in most situations; its empirical size is close

**Table 2.** Alcohol consumption data.

Age (Years)		Daily Alcohol Consumption		Odds Ratio
		80 + g	0 - 79 g	
25 - 34	Case	1	0	33.63
	Control	9	106	
35 - 44	Case	4	5	5.05
	Control	26	164	
45 - 54	Case	25	21	5.67
	Control	29	138	
55 - 64	Case	42	34	6.36
	Control	27	139	
65 - 74	Case	19	26	2.58
	Control	18	88	
75+	Case	5	8	40.76
	Control	0	31	

Source: Statistical Methods in Cancer Research, volume 1, page 137.

**Table 3.** Homogeneity odds ratio test results for alcohol consumption data.

Statistics	Observed value	Expected value	Variance	z value	p-value
U3	2.24525	0.452702	0.144733	2.24525	0.000001228
WU3	6.25017	4.47819	8.0402	0.624919	0.26601
Breslow_Day	9.38159				0.0968
Zelen exact	0.0968552				0.0969

**Table 4.** New drug data.

Test Site	New Drug		Control Drug		Odds Ratio
	Response	No	Response	No	
1	0	15	0	15	1
2	0	39	6	32	0.06
3	1	20	3	18	0.3
4	1	14	2	15	0.54
5	1	20	2	19	0.48
6	0	12	2	10	0.17

**Continued**

7	3	49	10	42	0.26
8	0	19	2	17	0.18
9	1	14	0	15	1.07
10	2	26	2	27	1.04
11	0	19	2	18	0.19
12	0	12	1	11	0.31
13	0	24	5	19	0.07
14	2	10	2	11	1.1
15	0	14	11	3	0.01
16	0	53	4	48	0.1
17	0	20	0	20	1
18	0	21	0	21	1
19	1	50	1	48	0.96
20	0	13	1	13	0.32
21	0	13	1	13	0.32
22	0	21	0	21	1

**Table 5.** Homogeneity odds ratio test results for new drug data.

Statistics	Observed value	Expected value	Variance	z value	p-value
U3	1.3056	0.762194	0.117405	1.58593	0.0564
WU3	6.56955	4.11497	1.09492	2.34578	0.0095
Breslow-Day	25.7844				0.0785
Zelen exact	0.0292015				0.0292

to 0.05 when the sample size is large; when sample size is small, Breslow-Day test is not recommended. Breslow-Day test is never recommended for sparse data;

When the sample size is small and the number of strata is small, say less than 5, Zelen's exact test is recommended;

In our application, the sample mean and the variance were estimated based on certain assumptions. The empirical power and size of  $U3$  and  $WU3$  would be highly dependent on how well the estimator of  $\sigma^2$  would be.

### Acknowledgements

This work was partially supported by Cancer Prevention Research Institute of Texas (RP170668).

### Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

- [1] Liang, K.Y. and Self, S.G. (1985) Tests of Homogeneity of Odds Ratio When the Data Are Sparse. *Biometrika*, **72**, 353-358. <https://doi.org/10.1093/biomet/72.2.353>
- [2] Paul, S.R. and Donner, A. (1992) Small Sample Performance of Tests of Homogeneity of Odds Ratios in  $k \times 2$  Tables. *Statistics in Medicine*, **11**, 159-165. <https://doi.org/10.1002/sim.4780110203>
- [3] Jones, M.P., O’Gorman, T.W., Lemke, J.H. and Woolson, R.F. (1989) A Monte Carlo Investigation of Homogeneity Tests of the Odds Ratio under Various Sample Size Configurations. *Biometrics*, **45**, 171-181. <https://doi.org/10.2307/2532043>
- [4] Zelen, M. (1971) The Analysis of Several  $2 \times 2$  Contingency Tables. *Biometrika*, **58**, 129-137. <https://doi.org/10.1093/biomet/58.1.129>
- [5] Halperin, M., Ware, J.H., Byar, D.P., Mantel, M., Brown, C.C., Koziol, J., Jail, M. and Green, S.B. (1977) Testing for Interaction in an  $I \times J \times K$  Contingency Table. *Biometrika*, **64**, 271-275. <https://doi.org/10.1093/biomet/64.2.271>
- [6] Thomas, D.G. (1975) Exact and Asymptotic Methods for the Combination of  $2 \times 2$  Tables. *Computers and Biomedical Research*, **8**, 423-446. [https://doi.org/10.1016/0010-4809\(75\)90048-8](https://doi.org/10.1016/0010-4809(75)90048-8)
- [7] Pagano, M. and Tritchler, D. (1983). Algorithms for the Analysis of Several  $2 \times 2$  Contingency Tables. *SIAM Journal of Scientific and Statistical Computing*, **4**, 302-309. <https://doi.org/10.1137/0904024>
- [8] Metha, C.R., Patel, N.R. and Wei, L.J. (1988) Computing Exact Significance Tests with Restricted Randomization Rules. *Biometrika*, **75**, 295-302. <https://doi.org/10.1093/biomet/75.2.295>
- [9] Thomas, D.G. and Gart, J.J. (1992) Improved and Extended Exact and Asymptotic Methods for the Combination of  $2 \times 2$  Tables. *Computers and Biometrical Research*, **25**, 75-84. [https://doi.org/10.1016/0010-4809\(92\)90036-A](https://doi.org/10.1016/0010-4809(92)90036-A)
- [10] Hirji, K.F., Vollset, S.E., Reis, I.M. and Afifi, A.A. (1996) Exact Tests for Interaction in Several  $2 \times 2$  Tables. *Journal of Computational and Graphical Statistics*, **5**, 209-224. <https://doi.org/10.1080/10618600.1996.10474706>
- [11] Reis, I. M., Hirji, K.F. and Afifi, A.A. (1999) Exact and Asymptotic Tests for Homogeneity in Several  $2 \times 2$  Tables. *Statistics in Medicine*, **18**, 893-906. [https://doi.org/10.1002/\(SICI\)1097-0258\(19990430\)18:8<893::AID-SIM84>3.0.CO;2-5](https://doi.org/10.1002/(SICI)1097-0258(19990430)18:8<893::AID-SIM84>3.0.CO;2-5)
- [12] Bagheri, Z., Ayatollahi, S.M.T. and Jafari, P. (2011) Comparison of Three Tests of Homogeneity of Odds Ratios in Multicenter Trials with Unequal Sample Sizes within and among Centers. *BMC Medical Research Methodology*, **11**, 58. <https://doi.org/10.1186/1471-2288-11-58>
- [13] Viwatwongkasem, C., Donjdee, K. and Poodphraw, T. (2018) Profile Likelihood Tests for Common Risk Ratios in Meta-Analysis Studies. *Open Journal of Statistics*, **8**, 915-930. <https://doi.org/10.4236/ojs.2018.86061>
- [14] Halmos, P.R. (1946) The Theory of Unbiased Estimation. *The Annals of Mathematical Statistics*, **17**, 34-43. <https://doi.org/10.1214/aoms/1177731020>
- [15] Hoeffding, W. (1948) A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, **19**, 293-325. <https://doi.org/10.1214/aoms/1177730196>
- [16] Serfling, R.J. (1980) Approximation Theorems of Mathematical Statistics. Wiley, New York. <https://doi.org/10.1002/9780470316481>
- [17] Lee, A.J. (1990) U-Statistics: Theory and Practice. Dekker, New York.

- [18] Gart, J.J. and Zweifel, J.R. (1967) On the Bias of Various Estimators of the Logit and Its Variance, with Application to Quantal Bioassay. *Biometrika*, **54**, 181-187.  
<https://doi.org/10.1093/biomet/54.1-2.181>
- [19] Breslow, N.E. and Day, N.E. (1980) Statistical Methods in Cancer Research. Vol. 1—The Analysis of Case-Control Studies. International Agency for Research on Cancer (IARC). Lyon, France, 136-146.
- [20] Mehta, C. and Patel, N. (2002) StatXact 4 for Windows User Manual. Cytel Software, Cambridge, MA, 486-490.