

Contingency Table Analysis in Obstetrics and Gynaecology

Daniel YT FONG PhD

Chun Fan LEE MPhil

Siu Pik LAU MPH

Department of Nursing Studies, The University of Hong Kong

Contingency table analysis for investigating the association between two categorical variables is common. Methods often employed are the χ^2 test and Fisher's exact test. However, they have been sometimes misused and there has been no systematic evaluation in obstetrics and gynaecology. Therefore, we aimed to describe the available methods of analysing contingency tables and to evaluate the performance of recent literature in obstetrics and gynaecology. Searching the first three issues of the *Obstetrics and Gynecology* journal in 2008 and 2003 identified respectively 34 (57%) and 46 (62%) studies with the use of χ^2 test, Fisher's exact test, or exact χ^2 test specified or actually performed. However, only 22 (in 2008) and 28 (in 2003) studies have sufficient data for verifying the actual use of tests; of which, 82% (exact 95% confidence interval = 60% to 95%) and 67% (exact 95% confidence interval = 46% to 83%) respectively had at least one inadequate use. Those common ones included non-reproducible p values, using χ^2 test without fulfilling its requirements, and using χ^2 test for 2x2 tables when Fisher's exact test is available. No significant differences between the two publication years were observed. Exact χ^2 test, though not widely implemented, is more appropriate for contingency table analysis. Substantial inadequate use of tests in contingency table analysis was observed in a highly ranked journal in obstetrics and gynaecology and no notable improvement if not worsening was observed over the past 5 years. More effective communications between clinicians and statisticians are required.

Hong Kong J Gynaecol Obstet Midwifery 2008; 8:42-50

Keywords: Chi-square distribution; Statistics

Introduction

Contingency table analysis for assessing the association between two categorical variables is common in all disciplines including obstetrics and gynaecology. For instance, Coutinho et al¹ reported more women with exteriorised repair surgery experienced moderate or severe pain 6 hours after surgery than women underwent in-situ repair. Qiu et al² compared singleton- and multiple-birth mothers in terms of their use of drugs, alcohols, tobacco, etc. In such instances, the chi-square (χ^2) test for association is the most frequently used method³. Besides, the Fisher's exact test has also been recommended when the sample size is small^{4,5}. Despite their vast popularity, there have been reported cases of their misuses, including non-reproducible p values and their use in paired samples⁵⁻¹⁰. Indeed, with the advances in statistics and computing resources, contingency table analysis may be performed more accurately by the exact χ^2 test. However, how the three tests are compared and when they are used have not been discussed in

the literature. Consequently, a better understanding of the methods is deemed to be necessary to ensure a contingency table analysis is properly performed.

Welch and Gabbe¹⁰ reviewed the general use of statistics in 145 papers published in the *American Journal of Obstetrics and Gynecology* in 1994. A total of 32% had inappropriate use of statistics but it was reduced to 10% in a subsequent review in 1997 conducted by the same group of researchers¹¹. However, there were no systematic assessments on contingency table analysis in the literature, particularly in obstetrics and gynaecology. Therefore, we aimed to describe and contrast the available methods of analysing contingency tables and to evaluate the performance of recent literature

Correspondence to: Dr DYT Fong, Department of Nursing Studies, The University of Hong Kong, Pokfulam Road, Hong Kong

Tel: (852) 2819 2645

Fax: (852) 2872 6079

Email: dytfong@hku.hk

Table 1. The 2x2 contingency table of the infrared gun (IRS) study

Pain improvement	Treatment		Total
	IRS	Placebo	
Yes	9	4	13
No	3	9	12
Total	12	13	25

in obstetrics and gynaecology.

Contingency Table Analysis

The IRS Study

To ease illustration, we consider a randomised, controlled trial conducted to compare the effectiveness of an infra-red gun (IRS Medtec 100) with a mock transcutaneous nerve (TNS; placebo) in treating patients with cervical osteoarthritis pain¹². A total of 25 patients with cervical osteoarthritis pain completed the study with 13 patients received mock TNS (placebo) and 12 patients underwent IRS. The corresponding 2x2 contingency table of results is shown in Table 1. The research question was whether IRS was different from placebo in treating cervical osteoarthritis pain. Here we are testing the null hypothesis of no association between treatment (IRS or placebo) and cervical osteoarthritis pain against the alternative hypothesis of an association. Therefore, a sufficiently small p value indicates the rejection of the null hypothesis and leads to the conclusion of significant association. Otherwise, there is not sufficient evidence to conclude association.

The χ^2 Test

The χ^2 test was developed by the very well-known mathematician, Karl Pearson, in July 1900 as a goodness-of-fit test^{3,13}. Therefore, it is also known as the Pearson χ^2 test. The χ^2 test is a non-parametric test which bears no assumption on the underlying population and thus can be applied without the worry of testing any distributional assumptions. For 2x2 tables, the χ^2 test can also be taken as testing about the difference between two proportions.

It is important to note that the p value of a χ^2 test is calculated asymptotically, i.e. based on the assumption that the sample size is sufficiently large. In other words, the χ^2 test is only an approximation but its accuracy increases with larger sample size. To conclude

if a sample is sufficiently large is often guided by the two conditions¹⁴:

1. All cells should have expected frequencies of ≥ 1 ; and
2. At least 80% of the cells in the contingency table have expected frequencies of ≥ 5 .

When the two required conditions are not fulfilled, the accuracy of χ^2 test may become questionable.

When the sample size is small, one sometimes uses the continuity correction suggested by Yates in 1934¹⁵. However, it is often criticised to be too conservative¹⁶⁻¹⁹. This can be illustrated by the IRS study where the uncorrected and corrected p values are 0.027 and 0.070, respectively. That is, the correction alters the conclusion from significance to insignificance if we operate at 5% level of significance. So, should we conclude a significant association or not?

Indeed, the use of Yates continuity correction remains controversial. While some recommended its use^{20,21}, Zar²² recommended using it only for 2x2 tables and some others even recommended not using it at all¹⁶⁻¹⁹. Nevertheless, the common statistical packages such as SPSS provided the corrected p value for 2x2 tables only. We are not extending nor resolving the controversy here but suggesting a better option.

The Exact χ^2 Test

The reliance of large sample size in the calculation of p value of the χ^2 test is fortunately not required with today's advances in statistical science and computing resources. Various algorithms have been developed to facilitate the calculation of exact p value without requiring a large sample size²³. However, the methods have not been widely implemented. To our knowledge, statistical packages that support the exact χ^2 test were StatXact, Statistical Analysis System (SAS) and SPSS Exact Tests²⁴⁻²⁶.

In the IRS study, the exact p value is 0.047. Therefore, the association between treatment and pain improvement should be taken as statistically significant at 5% level of significance.

The Fisher's Exact Test

The Fisher's exact test is the most commonly

used alternative to χ^2 test when the sample size is small. It was developed by a famous statistician Sir R.A. Fisher in 1925 for 2x2 tables and later generalised to tables of larger size^{27,28}. This test gives an exact p value and thus is also not restricted to large samples. However, it assumes the marginal, i.e. row and column, totals of the contingency table are fixed across all possible samples, an assumption which may rarely be true in practice⁵.

In 2x2 table analysis, the Fisher's exact test can be 1-sided or 2-sided. Note the examination of association in a 2x2 table can be equivalent to testing the difference between two proportions. A 2-sided test assesses the difference between the two proportions while a 1-sided test examines only one side of the difference, e.g. one proportion is larger than the other, while ignoring the other side of the difference. In the sequel, one should use the 1-sided test only when the occurrence of the other side of the difference is absolutely irrelevant²⁹. This is again in practice rarely happened but the sidedness of the test should nevertheless be specified especially when a 1-sided test is used⁵.

In general, the computation of p values by the Fisher's exact test increases drastically with the number of cells in the contingency table. Therefore, most statistical packages do not perform the test in tables larger than 2x2.

The Recommended Approach

The analysis of association between two categorical variables should preferably be performed by the exact χ^2 test. This gives the exact and most accurate p values without the worry of sample size.

However, one may not have a statistical package with this test implemented. Under such circumstance, we may just use the Fisher's exact test for 2x2 tables when the Fisher's exact test is equivalent to the exact χ^2 test³⁰. Therefore, although the Fisher's exact test is often recommended when the sample size is small, it can also be used otherwise to give the exact p value, as long as the statistical package permits. For tables other than 2x2, we prefer using χ^2 test since the Fisher's exact test relies on the unnatural assumption of fixed marginal totals which is rarely true in practice. However, we need to first check for the validity of the two required conditions

of χ^2 test. In case when the conditions are not fulfilled, we may try collapsing some neighbouring columns or rows until the required conditions are satisfied. In the worst case, the table may be reduced to 2x2 when the Fisher's exact test can be employed without fulfilling the required conditions.

Evaluation of Recent Literature in Obstetrics and Gynaecology

Identification of Literature

A hand search was made on the most recent issues of the *Obstetrics and Gynecology* journal published in the year of 2008, i.e. No. 1-3 of Volume 111. The *Obstetrics and Gynecology* journal had the largest circulation and ranking impact factor in all peer-reviewed obstetrics and gynaecology journals³¹. Therefore, assessing this journal would provide the most optimistic indication of performance in the obstetrics and gynaecology literature. Moreover, in order to assess the potential changes over the years, the first three issues of the journal in 2003 were also searched. Studies included were those that specified the use or actually used the χ^2 test, Fisher's exact test, or the exact χ^2 test in contingency table analysis.

Data relevant to the use of tests were extracted. Particularly, when there were sufficient data reported, the tests indicated in the studies were re-done to determine if the results were reproducible. Difference between the two publication years were compared by the exact χ^2 test for categorical variables and by exact Wilcoxon rank sum test for continuous variables. Paired comparisons of continuous variables were performed by signed rank test. A 5% level of significance was used and all analyses were performed by the SAS version 9.

Evaluation Results

A total of 60 and 74 studies were published in the first three issues of the *Obstetrics and Gynecology* journal in 2008 and 2003, respectively. Of which, 34 (57%) and 46 (62%) had at least one of the three tests for contingency table analysis specified in the methods section or actually performed. A summary is provided in Table 2. Only studies published in 2008 had the level of evidence since the journal only began this identification in 2004. The SAS, SPSS, and Stata were the most popularly used packages which constituted 65% of studies identified in 2008 and 50% in 2003.

Table 2. Identified studies that examined the association between two categorical variables

	2008 (n=34)*		2003 (n=46)*		Difference
	n	(%)	n	(%)	p value [†]
Level of evidence					
I	8	(23.5)	-		
II	19	(55.9)	-		
III	7	(20.6)	-		
Statistical packages used					0.272
Not mentioned	7	(20.6)	18	(39.1)	
EpiInfo	1	(2.9)	1	(2.2)	
SPSS	8	(23.5)	6	(13.0)	
SAS	8	(23.5)	12	(26.1)	
SAS/JMP	1	(2.9)	0	(0)	
Minitab	0	(0)	1	(2.2)	
R	1	(2.9)	0	(0)	
StatXact	0	(0)	2	(4.3)	
StatView	0	(0)	1	(2.2)	
Stata	6	(17.6)	5	(10.9)	
SUDAAN	1	(2.9)	0	(0)	
SigmaStat	1	(2.9)	0	(0)	
Tests indicated in methods section					0.027
χ^2 test only	22	(64.7)	19	(41.3)	
Fisher's exact test only	1	(2.9)	12	(26.1)	
Exact χ^2 test only	1	(2.9)	1	(2.2)	
χ^2 test and Fisher's exact test	10	(29.4)	14	(30.4)	
Tests actually used					0.115
Not verifiable due to insufficient data	12	(35.3)	19	(41.3)	
None	1	(2.9)	1	(2.2)	
χ^2 test only	10	(29.4)	8	(17.4)	
Fisher's exact test only	0	(0)	9	(19.6)	
Exact χ^2 test only	1	(2.9)	1	(2.2)	
χ^2 test and Fisher's exact test	9	(26.5)	6	(13.0)	
Fisher's exact test and exact χ^2 test	0	(0)	1	(2.2)	
χ^2 test, Fisher's exact test and exact χ^2 test	1	(2.9)	1	(2.2)	
Analysed a 2x2 table	24	(70.6)	42	(91.3)	0.020

* Identified by a hand search of articles published in the first three issues of the year

[†] By exact χ^2 test

There was only one study in the respective year that used the exact χ^2 test. This may probably due to the incapability of most statistical packages (Table 3). On the other hand, the tests described in the methods section may be inconsistent to the tests actually performed. The discrepancies are summarised in Table 4. Moreover, there were two studies in 2008 and one study in 2003

that inappropriately specified the use of χ^2 test or Fisher's exact test for testing about discrete variables, but the variables were actually categorical. There were respectively five (23%, exact 95% CI = 8-45%) of 22 verified and six (22%, exact 95% CI = 9-42%) of 27 verified studies in 2008 and 2003 that had inadequate description of tests (p=1.000).

Table 3. Availability of the Fisher’s exact test and the exact χ^2 test in statistical packages used in the identified studies

	Fisher’s exact test	Exact χ^2 test
EpiInfo (version 3.4.1)	✓ (only for 2x2 tables)	✗
SPSS Base (versions 14, 15)	✓ (only for 2x2 tables)	✗
SAS (versions 8, 9)	✓	✓
SAS/JMP (version 7)	✓ (only for 2x2 tables)	✗
Minitab (versions 14, 15)	✓ (only for 2x2 tables)	✗
R (version 2.6.2)	✓	✗
StatXact (version 5.8)	✓	✓
StatView (version 5.0)	✓ (only for 2x2 tables)	✗
Stata (version 9.2)	✓	✗
SUDAAN (version 9)	✗	✗
SigmaStat (version 3.1)	✓ (only for 2x2 tables when cell counts ≤ 5)	✗

Table 4. Differences between tests described in the methods section and the tests actually used

	2008			2003			Difference in the actual use of tests p value*
	Total	With sufficient data for verification n (%)	Actually used n (%)	Total	With sufficient data for verification n (%)	Actually used n (%)	
Test specified in the methods section†							
χ^2 test	32	20 (62.5)	19 (95.0)	33	15 (45.5)	15 (100)	1.000
Fisher’s exact test	12	8 (66.7)	8 (100)	26	17 (65.4)	15 (88.2)	1.000
Exact χ^2 test	1	1 (100)	1 (100)	1	1 (100)	1 (100)	1.000
Test not specified in the methods section†							
χ^2 test	2	2 (100)	1 (50.0)	13	10 (76.9)	0 (0)	0.167
Fisher’s exact test	22	19 (86.4)	2 (10.5)	20	10 (50.0)	2 (20.0)	0.592
Exact χ^2 test	33	21 (63.6)	1 (4.8)	45	24 (53.3)	2 (8.3)	1.000

* By exact χ^2 test

† Numbers under the column may not add up to the corresponding total since a study may use more than one test

Table 5 shows inadequate uses of tests in contingency table analysis. For the χ^2 test, 10 (50%, exact 95% CI = 27-73%) and eight (53%, exact 95% CI = 27-79%) verified studies in 2008 and 2003 respectively either had non-reproducible p values or did not have the two required conditions of χ^2 test satisfied. For the Fisher’s exact test, six (60%, exact 95% CI = 26-88%) verified studies in 2008 either had non-reproducible p values or 1-sided without specifying the direction of difference, and three (24%, exact 95% CI = 7-50%) verified studies in 2003 had non-reproducible p

values. We also noted three studies in 2003 and none in 2008 used Fisher’s exact test in tables other than 2x2. Moreover, all studies did not indicate if Fisher’s exact test was performed as 1-sided or 2-sided. For the analysis of 2x2 tables, 11 (65%, exact 95% CI = 38-86%) and 10 (43%, exact 95% CI = 23-66%) verified studies in 2008 and 2003 respectively had either used Fisher’s exact test only when the sample size was small or used χ^2 test only. On the other hand, there were no studies that used the tests in paired samples. Besides, there were no general differences in the various inadequate uses between the

Table 5. Inadequate uses of tests in contingency table analysis

	2008		2003		Difference
	n	(%)	n	(%)	p value*
χ^2 test	(n=20 verified)		(n=15 verified)		
Non-reproducible p values	9	(45.0)	5	(35.7)	0.728
Did not fulfill the requirements of χ^2 test	6	(30.0)	7	(46.7)	0.481
Fisher's exact test	(n=10 verified)		(n=17 verified)		
Non-reproducible p values	4	(40.0)	3	(17.6)	0.365
Used 1-sided test without defined direction of association	2	(20.0)	0	(0)	0.128
Analysis of 2x2 tables	(n=17 verified)		(n=23 verified)		
Used Fisher's exact test only for small sample size	4	(23.5)	4	(17.4)	0.702
Used χ^2 test only	7	(41.2)	6	(26.1)	0.496

* By exact χ^2 test

Table 6. Discrepancies of p values in studies with non-reproducible test results

	2008				2003				Difference
	n	Mean (SD)	Median (range)	p value*	n	Mean (SD)	Median (range)	p value*	p value†
χ^2 test									
Reported χ^2 test	9	-0.042 (0.257)	-0.008 (-0.539 to 0.451)	0.426	5	-0.072 (0.155)	-0.008 (-0.348 to 0.009)	0.625	0.925
Reported exact χ^2 test	9	-0.048 (0.268)	-0.001 (-0.577 to 0.452)	0.570	5	-0.128 (0.151)	-0.070 (-0.349 to 0.008)	0.125	0.364
Fisher's exact test									
Reported	4	-0.156 (0.411)	0.024 (-0.770 to 0.099)	0.813	3	0.012 (0.179)	-0.024 (-0.146 to 0.206)	0.625	1.000
Fisher's exact test									

* By signed rank test

† By exact Wilcoxon rank sum test

two publication years ($p > 0.05$). The discrepancies in p values for studies with non-reproducible p values are summarised in Table 6. There was no evidence that the p values were tended to be over- or under-reported.

Overall, 18 (82%, exact 95% CI = 60-95%) and 17 (63%, exact 95% CI = 42-81%) of the verified studies in 2008 and 2003 respectively had at least one inadequate use of a test (p for the difference between the two years = 0.207). The median number of inadequate use was 1 (range, 1-4) in both publication years. In 2008, the number of inadequate uses was 6 (out of 6 verified, 100%), 10 (out of 13 verified, 77%) and 2 (out of 3 verified, 67%) in studies with level of evidence I, II, and III respectively. There was no evidence on the association between inadequate use of tests and the level of evidence of the studies ($p = 0.414$).

Discussion

Contingency table analysis will continue to be common in practice, and its appropriateness and proper reporting of methods and results are essential. We described and contrasted two frequently used methods and an exact method. The χ^2 test is most frequently used but it is only an approximation though its accuracy increases with larger sample size. The Yates continuity correction would give conservative p values and its use is becoming obsolete. The Fisher's exact test bears the unnatural assumption that all marginal totals remain fixed across all possible samples. On the other hand, the exact χ^2 test is preferable as it is not restricted by small sample size and does not bear any unnatural assumptions. Nevertheless, the 2-sided Fisher's exact test gives p values identical to the exact χ^2 test for 2x2 tables.

In the highly ranked journal in the obstetrics and gynaecology literature, there was moderate proportion (18% in 2008 and 21% in 2003) of verified studies with inadequate description of tests. The tests described in the methods section and the tests actually used may not be consistent. Although such discrepancies were also previously reported in rehabilitation research, they should have been easily avoided if careful proofreading was performed³².

It is somewhat surprising to observe around 50% of studies verified did not use or report the χ^2 test adequately with either p values not reproducible or the test was used without fulfilling the required conditions on expected frequencies. Non-reproducible results have unfortunately been not uncommon and may also happen in major journals. A review conducted in 2001 revealed 11.6% (21/181) and 11.1% (7/63) of the statistical results published in *Nature* and *BMJ* journals respectively were not reproducible³³. This may be due to inadequate reporting of results. We speculate that some studies may not have reported the number of missing values which brought to the discrepancies in p values when tests were performed based on the reported data. Therefore, more careful reporting and checking of results is needed. On the other hand, most statistical packages do provide expected frequencies upon request. Particularly, common statistical packages such as SAS and SPSS routinely counts the number of cells with expected frequency less than five and may even give warning when either of the required conditions is violated. However, it seems there has been an ignorance of checking the requirements.

There was a small proportion (20% in 2008 and none in 2003) of studies in the *Obstetrics and Gynecology* journal that used a 1-sided Fisher's exact test without any specification in the methods section. Indeed, 36% of 56 selected studies in six top medical journals between 1983 and 1987 also used the 1-sided Fisher's exact test without specifying the sidedness⁵. The 1-sided Fisher's exact test is rarely used in practice since we often cannot rule out the occurrence of any side of the difference. Nevertheless, using the test without indicating the sidedness may mislead the interpretation of statistical significance since the p value of a 1-sided test is always

smaller than the 2-sided version⁵. Particularly, under no circumstances should one use a 1-sided test because of a significant p value.

In the analysis of 2x2 tables, 65% of the recent literature verified (43% in 2003) used the χ^2 test. The analysis could have been improved by using the Fisher's exact test which is equivalent to the exact χ^2 test for the analysis of 2x2 tables. The Fisher's exact test has often been recommended when the sample size is small but it appears no reason why it cannot be used in larger samples with today's advances in computing resources. Indeed, the test has been implemented, at least for 2x2 tables, in many statistical packages or freely available online. It is therefore highly accessible and the consistent use of it, as an equivalent to the exact χ^2 test, is recommended irrespective to the sample size.

The substantial inadequacy (82% in 2008 and 63% in 2003) in contingency table analysis in obstetrics and gynaecology research may probably be due to insufficient collaborations or communications between clinicians and statisticians. Statisticians need to have reasonable knowledge and understanding of the clinical contents and problems before embarking on the statistical analysis. Indeed, statisticians should be involved in the whole research process from design to reporting and interpretation of results, rather than merely performing the data analysis. All inadequacies we identified were indeed avoidable should there be good rapport between clinicians and statisticians. Both of them must learn to communicate more effectively and to be willing to collaborate with each other³⁴. Indeed, training on communication and collaboration has been recommended in future curriculum in biostatistics³⁵.

Nevertheless, we did not observe a worsening nor improvement on inadequate contingency table analysis in the highly ranked journal in obstetrics and gynaecology. This may certainly be due to insufficient sample size but there was likely no improvement if not worsened over the past 5 years. Real efforts should be made to provide reliable and valid statistical results by effective communications and collaborations between clinicians and statisticians.

References

1. Coutinho IC, Ramos de Amorim MM, Katz L, et al. Uterine exteriorization compared with in situ repair at cesarean delivery: a randomized controlled trial. *Obstet Gynecol* 2008; 111:639-47.
2. Qiu X, Lee SK, Tan K, Piedboeuf B, Canning R; Canadian Neonatal Network. Comparison of singleton and multiple-birth outcomes of infants born at or before 32 weeks of gestation. *Obstet Gynecol* 2008; 111:365-71. Erratum in: *Obstet Gynecol* 2008; 111:1217.
3. Plackett RL. Pearson Karl and the Chi-squared test. *Int Stat Rev* 1983; 51:59-72.
4. Colton T. Statistics in medicine. *Boston: Little Brown*, 1974, pp xii, 372.
5. McKinney WP, Young MJ, Hartz A, et al. The inexact use of Fisher's Exact Test in six major medical journals. *JAMA* 1989; 261:3430-3.
6. Dijkers MP. Misuse of the Pearson chi-square test of association. *Arch Phys Med Rehabil* 2005; 86:602; author reply 602.
7. Hoffman JI. The incorrect use of Chi-square analysis for paired data. *Clin Exp Immunol* 1976; 24:227-9.
8. Lewis D, Burke CJ. The use and misuse of the chi-square test. *Psychol Bull* 1949; 46:433-89.
9. Ottenbacher KJ. The chi-square test: its use in rehabilitation research. *Arch Phys Med Rehabil* 1995; 76:678-81.
10. Welch GE 2nd, Gabbe SG. Review of statistics usage in the American Journal of Obstetrics and Gynecology. *Am J Obstet Gynecol* 1996; 175:1138-41.
11. Welch GE 2nd, Gabbe SG. Statistics usage in the American Journal of Obstetrics and Gynecology: has anything changed? *Am J Obstet Gynecol* 2002; 186:584-6; 339.
12. Lewith GT, Machin D. A randomised trial to evaluate the effect of infra-red stimulation of local trigger points, versus placebo, on the pain caused by cervical osteoarthritis. *Acupunct Electrother Res* 1981; 6:277-84.
13. Berntson DA, et al. History of statistics & probability. [cited 2008 April 20]; Available from: <http://www.morris.umn.edu/~sungurea/introstat/history/w98/Pearson.html>.
14. Yates DS, Moore DS, Starnes DS. The practice of statistics: TI-83/89 graphing calculator enhanced. 2nd ed. *New York: WH Freeman*, 2003.
15. Yates F. Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society* 1936; Supplement 1:217-35.
16. Conover WJ. Uses and misuses of continuity correction. *Biometrics* 1968; 24:1028.
17. Conover WJ. Some reasons for not using Yates continuity correction on 2x2 contingency-tables. *J Am Stat Assoc* 1974; 69:374-6.
18. Haviland MG. Yates's correction for continuity and the analysis of 2 x 2 contingency tables. *Stat Med* 1990; 9:363-7; 369-83.
19. Grizzle JE. Continuity correction in Chi-2-Test for 2by2 tables. *Am Stat* 1967; 21:28-32.
20. Kirkwood BR, Sterne JA. Essential medical statistics. 2nd ed. *Malden, Mass: Blackwell Science*, 2003, pp x, 501.
21. Bland M. An introduction to medical statistics. 3rd ed. Oxford medical publications. *Oxford, UK: Oxford University Press*, 2000, pp xvi, 405.
22. Zar JH. Biostatistical analysis. 4th ed. *Upper Saddle River, NJ: Prentice Hall*, 1999, pp 1 v.
23. Agresti A. A survey of exact inference for contingency tables. *Stat Sci* 1992; 7:131-77.
24. Mehta CR, Patel NR. SPSS exact tests 7.0 for Windows. *Chicago, Ill: SPSS*, 1996, pp xiii, 220.
25. SAS Institute. SAS/STAT user's guide. 8th ed. *Cary, NC: SAS Institute*, 1999.
26. Mehta CR. Statxact - a Statistical Package for Exact Nonparametric-Inference. *Journal of Classification* 1990; 7:111-4.
27. Fisher RA. Statistical methods for research workers. Biological monographs and manuals. No. 5. *Edinburgh: Oliver and Boyd*, 1925, p239.
28. Freeman GH, Halton JH. Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* 1951; 38:141-9.
29. Cloft HJ, Kallmes DF. Statistical error in interpretation of aneurysm coil results. *Stroke* 2008; 39:e29; author reply e30.
30. Camilli G. The Relationship between Fishers Exact Test and Pearsons Chi-Square Test - a Bayesian Perspective. *Psychometrika* 1995; 60:305-12.
31. ISI Web of Knowledge. The Thompson Corporation. 2008.
32. Wainapel SF, Kayne HL. Statistical methods in

- rehabilitation research. *Arch Phys Med Rehabil* 1985; 66:322-4.
33. García-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. *BMC Med Res Methodol* 2004; 4:13.
34. Cheung YB, Tan SB, Khoo KS. The need for collaboration between clinicians and statisticians: some experience and examples. *Ann Acad Med Singapore* 2001; 30:552-5.
35. DeMets DL, Stormo G, Boehnke M, et al. Training of the next generation of biostatisticians: a call to action in the U.S. *Stat Med* 2006; 25:3415-29.