Fagerland *et al. BMC Medical Research Methodology* 2013, **13**:91
http://www.biomedcentral.com/1471-2288/13/91

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

The McNemar test for binary matched-pairs data: mid-*p* and asymptotic are better than exact conditional

Morten W Fagerland[1*], Stian Lydersen[2] and Petter Laake[3]

Abstract

Background: Statistical methods that use the mid-*p* approach are useful tools to analyze categorical data, particularly for small and moderate sample sizes. Mid-*p* tests strike a balance between overly conservative exact methods and asymptotic methods that frequently violate the nominal level. Here, we examine a mid-*p* version of the McNemar exact conditional test for the analysis of paired binomial proportions.

Methods: We compare the type I error rates and power of the mid-*p* test with those of the asymptotic McNemar test (with and without continuity correction), the McNemar exact conditional test, and an exact unconditional test using complete enumeration. We show how the mid-*p* test can be calculated using eight standard software packages, including Excel.

Results: The mid-*p* test performs well compared with the asymptotic, asymptotic with continuity correction, and exact conditional tests, and almost as good as the vastly more complex exact unconditional test. Even though the mid-*p* test does not guarantee preservation of the significance level, it did not violate the nominal level in any of the 9595 scenarios considered in this article. It was almost as powerful as the asymptotic test. The exact conditional test and the asymptotic test with continuity correction did not perform well for any of the considered scenarios.

Conclusions: The easy-to-calculate mid-*p* test is an excellent alternative to the complex exact unconditional test. Both can be recommended for use in any situation. We also recommend the asymptotic test if small but frequent violations of the nominal level is acceptable.

Keywords: Matched pairs, Dependent proportions, Paired proportions, Quasi-exact

Background

Matched-pairs data arise from study designs such as matched and crossover clinical trials, matched cohort studies, and matched case-control studies. The statistical analysis of matched-pairs studies must make allowance for the dependency in the data introduced by the matching. A simple and frequently used test for binary matched-pairs data is the McNemar test. Several versions of this test exist, including the asymptotic and exact (conditional) tests. The traditional advice is to use the asymptotic test in large samples and the exact test in small samples. The argument for using the exact test is that the asymptotic test may violate the nominal significance level for small sample sizes because the required asymptotics do not hold. One disadvantage with the exact test is conservatism: it produces unnecessary large *p*-values and has poor power.

Consider the data in Table 1, which gives the results from a study by Bentur et al. [1]. Airway hyperresponsiveness (AHR) status—an indication of pulmonary complications—was measured in 21 children before and after stem cell transplantation (SCT). The incidence of AHR increased from two (9.5%) children before SCT to eight (38%) children after SCT. The asymptotic test gives $p = 0.034$, and the exact test gives $p = 0.070$. The two *p*-values are considerably different, which often happens when we have a small sample size. The next example shows that this may also be the case for large sample sizes.

*Correspondence: morten.fagerland@medisin.uio.no
[1] Unit of Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway
Full list of author information is available at the end of the article

Table 1 Airway hyper-responsiveness (AHR) status before and after stem cell transplantation (SCT) in 21 children [1]

		After SCT		
		AHR	No AHR	Sum
Before SCT	AHR	1	1	2
	No AHR	7	12	19
	Sum	8	13	21

In another study of SCT, 161 myeloma patients received consolidation therapy three months after SCT [2]. Complete response (CR) was measured before and after consolidation (Table 2). An increase in CR following consolidation was observed: sixty-five (40%) patients had CR before consolidation compared with 75 (47%) patients after consolidation. The asymptotic test gives $p = 0.033$, and the exact test gives $p = 0.053$.

The choice between an asymptotic method and a conservative exact method—which can be summarized as a trade-off between power and preservation of the significance level—is well known from other situations involving proportions [3]. For the independent 2×2 table, a good compromise can be reached using the mid-p approach [4]. The Fisher mid-p test, which is a modification of Fisher's exact test, combines excellent power with rare and minor violations of the significance level [5]. The modification required to transform an exact p-value to a mid-p-value is simple: the mid-p-value equals the exact p-value minus half the point probability of the observed test statistic.

The purpose of this article is to investigate whether a mid-p version of the McNemar exact conditional test can offer a similar improvement for the comparison of matched pairs as has been observed with independent proportions. A supplementary materials document (Additional file 1) shows how the mid-p test can be calculated using several standard software packages, including Excel, SAS, SPSS, and Stata.

Methods

Notation

Let N denote the observed number of matched pairs of binomial events A and B—where the possible outcomes are referred to as success (1) or failure (2)—and let (Y_{i1}, Y_{i2}) denote the outcome of the ith pair. The observed data may be summarized in a 2×2 contingency table, as in Table 3. Each n_{kl} for $k, l = 1, 2$ corresponds to the number of event pairs (Y_{i1}, Y_{i2}) with outcomes $Y_{i1} = k$ and $Y_{i2} = l$. Let p_{kl} denote the joint probability that $Y_{i1} = k$ and $Y_{i2} = l$, which we assume independent of i. Following the notation in Agresti [6, pp. 418–420], this is a marginal or a population-averaged model. We denote the probabilities of success for events A and B—or equivalently, the marginal probabilities that $Y_{i1} = 1$ and $Y_{i2} = 1$—by p_{1+} and p_{+1}, respectively. The null hypothesis of interest is H_0: $p_{1+} = p_{+1}$. The alternative hypothesis is H_1: $p_{1+} \neq p_{+1}$.

It might, however, be more realistic to assume that p_{kl} also depends on the subject i. As denoted by Agresti [6, pp. 418–420], this is a subject-specific model. Further, this is a conditional model, since we are interested in the association within the pair, conditioned on the subject. Data from N matched pairs are then presented in N 2×2 tables, one for each pair. Collapsing over the pairs results in Table 3. Conditional independence between Y_1 and Y_2 is tested by the Mantel-Haenszel statistic [6, p.417]. But that test statistic is algebraically equal to the squared McNemar test statistic. In the following, we will not specify whether we test for marginal homogeneity or conditional independence.

The asymptotic McNemar test

The asymptotic McNemar test conditions on the number of discordant pairs $(n_{12} + n_{21})$. Conditionally, n_{12} is binomially distributed with parameters $n = n_{12} + n_{21}$ and $p = 1/2$ under the null hypothesis. The asymptotic McNemar test statistic [7], which is the score statistic for testing marginal homogeneity, is

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}, \tag{1}$$

and its asymptotic distribution is the standard normal distribution. The equivalent McNemar test statistic $\chi^2 = z^2 = (n_{12} - n_{21})^2/(n_{12} + n_{21})$ is approximately chi-squared distributed with one degree of freedom under the null hypothesis. The asymptotic McNemar test is undefined when $n_{12} = n_{21} = 0$.

Table 2 Complete response (CR) before and after consolidation therapy [2]

		After consolidation		
		CR	No CR	Sum
Before consolidation	CR	59	6	65
	No CR	16	80	96
	Sum	75	86	161

Table 3 The observed counts (and joint outcome probabilities) of a paired 2×2 table

		Event B		
		Success	Failure	Sum
Event A	Success	n_{11} (p_{11})	n_{12} (p_{12})	n_{1+} (p_{1+})
	Failure	n_{21} (p_{21})	n_{22} (p_{22})	n_{2+} (p_{2+})
	Sum	n_{+1} (p_{+1})	n_{+2} (p_{+2})	N (1)

The asymptotic McNemar test with continuity correction

Edwards [8] proposed the following continuity corrected version of the asymptotic McNemar test:

$$z = \frac{|n_{12} - n_{21}| - 1}{\sqrt{n_{12} + n_{21}}}. \quad (2)$$

The asymptotic McNemar test with continuity correction (CC) approximates the exact conditional test. Hence, it combines the disadvantage of an asymptotic test (significance level violations) with the disadvantage of a conditional exact test (overly conservativeness), and we do not expect it to perform well. We include it in our evaluations because it features in influential textbooks such as Altman [9] and Fleiss et al. [10]. The asymptotic McNemar test with continuity correction is undefined when $n_{12} = n_{21} = 0$.

The McNemar exact conditional test

The test statistic in (1) measures the strength of the evidence against the null hypothesis. If we, as in the derivation of the asymptotic test, condition on the number of discordant pairs ($n_{12} + n_{21}$), we can use the simple test statistic n_{12} to derive an exact conditional test. The conditional probability under H_0 of observing any outcome x_{12} given $n = n_{12} + n_{21}$ discordant pairs is the binomial point probability

$$f(x_{12}|n) = \binom{n}{x_{12}} \left(\frac{1}{2}\right)^n. \quad (3)$$

The McNemar exact conditional one-sided p-value is obtained as a sum of probabilities:

$$\text{one-sided } p\text{-value} = \sum_{x_{12}=0}^{\min(n_{12},n_{21})} f(x_{12}|n), \quad (4)$$

and the two-sided p-value equals twice the one-sided p-value. If $n_{12} = (n_{12} + n_{21})/2$, the p-value equals 1.0. The exact conditional test is guaranteed to have type I error rates not exceeding the nominal level.

The McNemar mid-p test

A mid-p-value is obtained by first subtracting half the point probability of the observed n_{12} from the exact one-sided p-value, then double it to obtain the two-sided mid-p-value [4]. Hence, the McNemar mid-p-value equals

$$\begin{aligned} \text{mid-}p\text{-value} &= 2 \cdot \left[\text{one-sided } p\text{-value} - \frac{1}{2} f(n_{12}|n)\right] \\ &= \text{two-sided } p\text{-value} - f(n_{12}|n), \end{aligned} \quad (5)$$

where f is the probability function in (3). If $n_{12} = n_{21}$, substitute (5) with

$$\text{mid-}p\text{-value} = 1 - \frac{1}{2} f(n_{12}|n). \quad (6)$$

The type I error rates of the mid-p test—as opposed to those of exact tests—are not bounded by the nominal level; however, in a wide range of designs and models, both mid-p tests and confidence intervals violate the nominal level rarely and with low degrees of infringement [11-13]. Because mid-p tests are based on exact distributions, they are sometimes called quasi-exact [14]. Additional file 1 provides details on how to calculate the McNemar mid-p test with several standard software packages.

An exact unconditional test

The tests in the previous sections did not used the concordant pairs of observations (n_{11} and n_{22}) in their calculations. The unconditional approach is to consider all possible tables with N pairs and thereby use information from all observed pairs, including the concordant ones. The exact unconditional test attributed to Suissa and Shuster [15] uses the McNemar test statistic (1). Let z_{obs} be the observed value, and let

$$z(\mathbf{x}) = \frac{x_{12} - x_{21}}{\sqrt{x_{12} + x_{21}}}, \quad (7)$$

where $\mathbf{x} = (x_{11}, x_{12}, x_{21}, x_{22})$ denotes a possible outcome with N pairs, and let $n = x_{12} + x_{21}$. If, for a one-sided test, $z_{\text{obs}} \geq 0$, the potential outcomes that provide at least as much evidence against the null hypothesis as the observed outcome—namely those with $z(\mathbf{x}) \geq z_{\text{obs}}$—are the pairs (x_{12}, n) in the region

$$C = \{(x_{12}, n) : x_{12} \geq h(n);\ x_{12} = 0, 1, \ldots, n;\ n = 0, 1, \ldots, N\}, \quad (8)$$

where $h(n) = 0.5 \cdot (z_{\text{obs}} n^{1/2} + n)$. Under the null hypothesis, the triplets $(x_{12}, n, N - n)$ are trinomially distributed with parameters N and $(p/2, p/2, 1 - p)$, and the attained significance level is

$$P(p) = \sum_C \binom{N}{x_{12}\ \ n - x_{12}\ \ N - n} \left(\frac{p}{2}\right)^n (1 - p)^{N-n}, \quad (9)$$

where p is the probability of a discordant pair (a nuisance parameter). We eliminate the nuisance parameter by maximizing $P(p)$ over the range of p. After simplifying (9), we get the following expression for the exact unconditional one-sided p-value [15]:

$$\text{one-sided } p\text{-value} = \sup_{0<p<1} \left[\sum_{n=k}^{N} \binom{N}{n} p^n (1 - p)^{N-n} F_n(n - i_n - 1)\right], \quad (10)$$

where $k = \text{int}(z_{\text{obs}}^2 + 1)$, F_n is the cumulative binomial distribution function with parameters $(n, 1/2)$, $i_n =$

int{$h(n)$}, and int is the integer function. Suissa and Shuster [15] outline a numerical algorithm to find the supremum in (10). If $z_{\text{obs}} < 0$, the one-sided p-value is found by reversing the inequality in (8). The two-sided p-value equals twice the one-sided p-value.

Evaluation of the tests

To compare the performances of the five tests, we carried out an evaluation study of type I error rates and power. We used complete enumeration (rather than stochastic simulations) and a large set of scenarios. Each scenario is characterized by fixed values of N (the number of matched pairs), p_{1+} and p_{+1} (the probabilities of success for each event), and $\theta = p_{11}p_{22}/p_{12}p_{21}$. θ can be interpreted as the ratio of the odds for the event Y_2 given Y_1. We use θ as a convenient way to re-parameterize $\{p_{11}, p_{12}, p_{21}, p_{22}\}$ into $\{p_{1+}, p_{+1}, \theta\}$, which includes the parameter of interest, namely the two marginal success probabilities. We used StatXact PROCs for SAS (Cytel Inc.) to calculate p-values of the exact unconditional test and Matlab R2011b (Mathworks Inc.) to calculate p-values of the four other tests and to perform the evaluation study. In cases where $n_{12} = n_{21} = 0$, we set $p = 1$ for the two asymptotic McNemar tests.

For the calculations of type I error rates, we used 19 values of N (10, 15, 20, …, 100), five values of θ (1.0, 2.0, 3.0, 5.0, 10.0), and 101 values of $p_{1+} = p_{+1}$ (0.00, 0.01, 0.02, …, 1.00), a total of 9595 scenarios. The nominal significance level was 5%.

Power was calculated for $N = 1, 2, \ldots, 100$, $\theta = 1.0, 2.0, 3.0, 5.0, 10.0$, $p_{1+} = 0.1, 0.35, 0.6$, and $\Delta = p_{+1} - p_{1+} = 0.10, 0.15, 0.20, 0.25, 0.30, 0.35$.

Results

Type I error rates

The between tests differences in type I error rates were largely consistent across the considered scenarios. Figure 1 illustrates these differences. The type I error rates of the McNemar exact conditional test are low and barely above 3%, even for as much as 100 matched pairs. The asymptotic McNemar test with CC performs similarly to the exact conditional test but is even more conservative. The asymptotic McNemar test (without CC) has type I error rates close to the nominal level for most combinations of parameters. It violates the level quite often, although not by much. The exact unconditional and the McNemar mid-p tests perform similarly. For most combinations of parameters, the type I error rates of the two tests are identical. For some situations with small proportions, however, the exact unconditional test has type I error rates closer to the nominal level than does the mid-p test (Figure 1, upper right and lower left panels). On the other hand, the mid-p test sometimes has type I error rates closer to the nominal level than does the exact

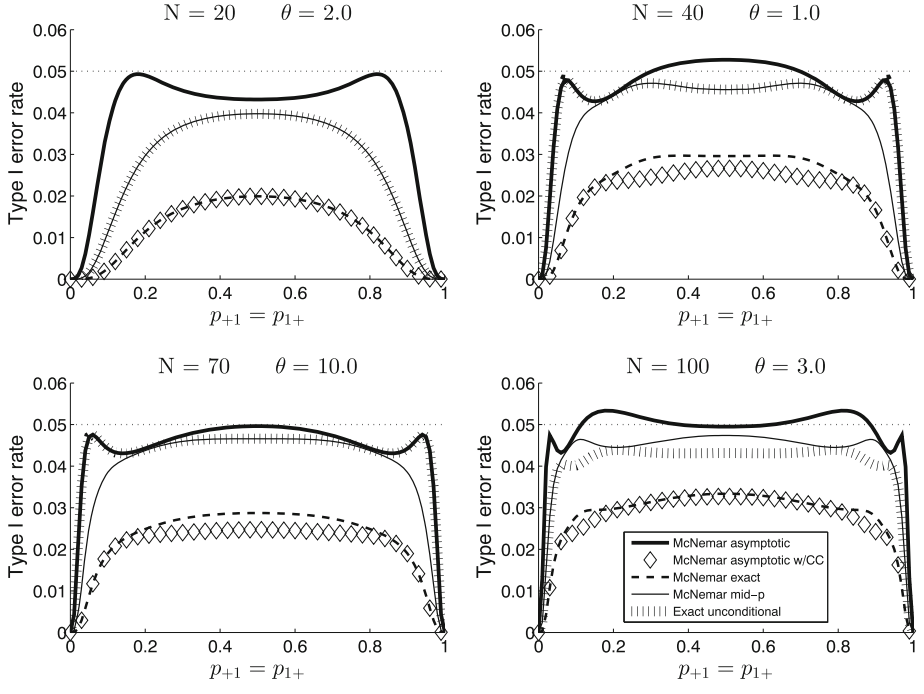

Figure 1 Type I error rates as functions of the probability of success. In each panel, the number of matched pairs (N) and the odds ratio (θ) are fixed.

unconditional test (Figure 1, lower right panel). Both tests are clearly superior to the McNemar exact conditional test and the asymptotic McNemar test with CC.

Table 4 presents summary statistics of the calculations of type I error rates. The mean and maximum type I error rate are shown for each test over all scenarios and for subregions based on the number of matched pairs. We also show the proportion of scenarios where the nominal significance level is violated and the proportion of scenarios where the type I error rate is below 3%. The asymptotic McNemar test violates the nominal significance level in 29% of the total number of considered scenarios. We note that this proportion is only 3.7% for small sample sizes ($10 \leq N \leq 30$) and as much as 52% for large sample sizes ($65 \leq N \leq 100$). A mitigating feature is that—as indicated in Figure 1—the infringement on the nominal significance level is small: the maximum type I error rate of the asymptotic McNemar test is 5.37%. If we are concerned with aligning the mean (instead of the maximum) type I error rate with the nominal level, the results in Table 4 suggest that the asymptotic McNemar test is the superior test, both overall and in each of the subregions based on sample size.

As expected, the two exact tests do not violate the nominal significance level in any of the considered scenarios. Interestingly, neither does the McNemar mid-p test.

Finally, one important comment to the interpretation of Table 4. The values of the parameters p_{1+} and p_{+1} were selected to represent the entire range of possible values and not to be a representative sample of the situations that might be encountered in practice. Scenarios with probabilities close to zero or one are thereby given more weight to the summary statistics in Table 4 than their impact in actual studies. Thus, the mean type I error rates of a typical study are likely closer to the nominal level than indicated in Table 4. The table is, however, a good illustration of the differences in performance between the five tests.

Further details of the results from the evaluation of type I error rates can be found in a supplementary materials document (Additional file 2), which contains box-plots of

Table 4 Evaluation of type I error rates (TIER)

Method	mean TIER	max TIER	proportion TIER > 0.05	proportion TIER < 0.03
All 9595 scenarios				
McNemar asymptotic	0.0430	0.0537	0.294	0.121
McNemar asymptotic w/CC	0.0190	0.0357	0.000	0.889
McNemar exact	0.0201	0.0367	0.000	0.880
McNemar mid-p	0.0349	0.0495	0.000	0.260
Exact unconditional	0.0373	0.0495	0.000	0.201
Subregion: $10 \leq N \leq 30$ (2525 scenarios)				
McNemar asymptotic	0.0352	0.0529	0.037	0.281
McNemar asymptotic w/CC	0.0089	0.0237	0.000	1.000
McNemar exact	0.0090	0.0278	0.000	1.000
McNemar mid-p	0.0212	0.0469	0.000	0.627
Exact unconditional	0.0251	0.0488	0.000	0.541
Subregion: $35 \leq N \leq 60$ (3030 scenarios)				
McNemar asymptotic	0.0435	0.0537	0.210	0.084
McNemar asymptotic w/CC	0.0196	0.0306	0.000	0.991
McNemar exact	0.0210	0.0306	0.000	0.989
McNemar mid-p	0.0374	0.0474	0.000	0.176
Exact unconditional	0.0408	0.0482	0.000	0.096
Subregion: $65 \leq N \leq 100$ (4040 scenarios)				
McNemar asymptotic	0.0476	0.0535	0.519	0.049
McNemar asymptotic w/CC	0.0249	0.0357	0.000	0.743
McNemar exact	0.0263	0.0367	0.000	0.723
McNemar mid-p	0.0416	0.0495	0.000	0.095
Exact unconditional	0.0423	0.0495	0.000	0.066

type I error rates from the total and various subregions of the evaluation study.

Power

Figure 2 shows the power of the tests as functions of the number of matched pairs with the usual yardsticks of 80% and 90% power marked in for reference. Only one combination of p_{1+}, p_{+1}, and θ is shown, however, the results where qualitatively equal for other settings. The powers of the asymptotic McNemar, the McNemar mid-p, and the exact unconditional tests are quite similar, although the asymptotic test is slightly better than the other two tests. The powers of the exact conditional test and the asymptotic McNemar test with CC trail that of the other tests considerably.

Table 5 displays the number of matched pairs needed to reach power of 50%, 60%, 70%, 80%, and 90% averaged over the 15 combinations of $\theta = 1.0, 2.0, 3.0, 5.0, 10.0$ and $p_{1+} = 0.1, 0.35, 0.60$. We show results for three of the Δ-values and note that similar results were obtained with $\Delta = 0.1$, 0.2, and 0.3. Values of N greater than 100 were estimated by simple linear extrapolation. The increase in sample size of using the exact unconditional or the mid-p test instead of the asymptotic McNemar test is quite small and in the range 0–3. The exact conditional test and the asymptotic McNemar test with CC, on the other hand, need a considerably greater sample size than the other tests. We emphasize that Table 5 is averaged over several combinations of parameters, and the values in it should not be used to plan the sample size of a study. The power of the tests are heavily dependent on the parameter values, even though the between tests differences in power were consistent across the different parameters in this evaluation. Table 5 thus illustrates typical sample size differences of the tests and not the actual sample size needed for a study.

The examples revisited

Table 6 presents the results of applying the five tests to the two examples introduced in the Background section. We have already observed that the asymptotic test and the exact conditional tests give quite different results for both examples. The asymptotic test with CC has p-values that are similar, but slightly higher, than the exact conditional test. The mid-p test and the exact unconditional test give results that largely agree with that of the asymptotic test. In both examples, the asymptotic, mid-p, and exact unconditional tests indicate stronger associations between airway hyper-responsiveness status and stem cell transplantation (Bentur et al. [1]) and between consolidation therapy and complete response (Cavo et al. [2]) than do the asymptotic test with CC and the exact conditional test. This difference in results is, perhaps, sufficiently great that different conclusions might be drawn. Because the asymptotic test with CC and the exact conditional test are highly conservative and have poor power, we do not recommend reporting the results of these two tests in any situation.

Discussion

The evaluation study in this article revealed several interesting observations. First, that the conservatism of the McNemar exact conditional test can be severe. A large sample size is needed to bring its type I error rates above 3% for a 5% nominal significance level. Quite often, the type I error rates of the exact conditional test were half that of the nominal level or lower. A similar conservative behavior has been observed for other exact conditional methods, for instance, Fisher's exact test for two independent binomial proportions [5] and the Cornfield exact confidence interval for the independent odds ratio [16]. This conservatism leads to poor power and a need for unnecessary large sample sizes. We do not

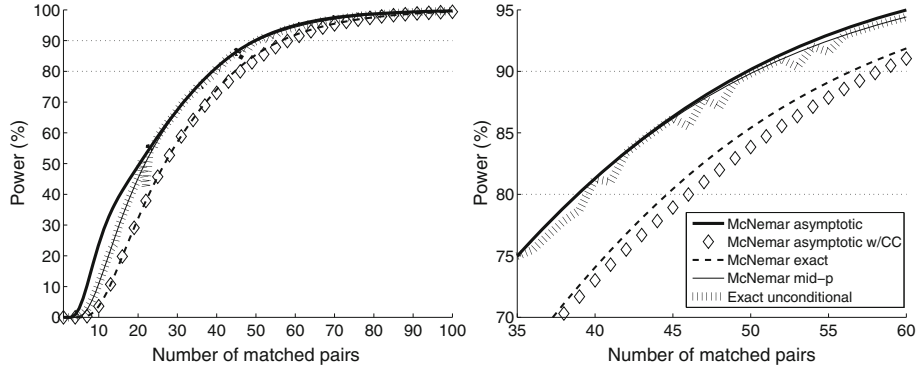

Figure 2 Power of the tests as functions of the number of matched pairs. The success probabilities ($p_{1+} = 0.1$ and $p_{+1} = 0.35$) and the odds ratio ($\theta = 2.0$) are fixed. The plot in the right panel shows details from the plot in the left panel.

Table 5 The number of matched-pairs (N) needed to reach power of 50%, 60%, 70%, 80%, and 90%, averaged over five values of θ and three values of p_{1+}, for three values of $\Delta = p_{+1} - p_{1+}$

	N to reach power of				
	50%	**60%**	**70%**	**80%**	**90%**
$\Delta = 0.15$					
McNemar asymptotic	56	69	85	103*	123*
McNemar asymptotic w/CC	68	82	98	114*	129*
McNemar exact	67	81	96	113*	129*
McNemar mid-*p*	57	71	87	104*	123*
Exact unconditional	57	71	88	106*	126*
$\Delta = 0.25$					
McNemar asymptotic	23	28	34	42	54
McNemar asymptotic w/CC	30	35	41	50	63
McNemar exact	29	34	41	49	62
McNemar mid-*p*	24	29	35	43	56
Exact unconditional	24	29	35	43	56
$\Delta = 0.35$					
McNemar asymptotic	13	15	18	23	29
McNemar asymptotic w/CC	18	21	24	28	34
McNemar exact	18	21	23	27	34
McNemar mid-*p*	15	17	20	24	30
Exact unconditional	15	17	20	24	29

*Values above 100 estimated by linear extrapolation.
This table should not be used for sample size calculations.

recommend use of the McNemar exact conditional test in any situation.

Second, the McNemar mid-p test is a considerable improvement over the exact conditional test on which it is based. It performs almost at the same level as the exact unconditional test. Whereas the exact tests are guaranteed to have type I error rates bounded by the nominal level, no such claim can be made for the mid-p test. Nevertheless, the mid-p test did not violate the nominal level in any of the 9595 scenarios considered in this evaluation. For practical use, the mid-p test is at an advantage vis-a-vis the exact unconditional test. As shown in the supplementary materials, the mid-p test is readily calculated in many commonly used software packages, including the ubiquitous Excel. The exact unconditional test, on the other hand, is computationally complex and only available in StatXact (Cytel Inc.).

Table 6 P-values of five tests using data from two published studies

	Bentur et al. [1]	**Cavo et al. [2]**
	2/21 vs 8/21	**65/161 vs 75/161**
McNemar asymptotic	0.0339	0.0330
McNemar asymptotic w/CC	0.0771	0.0550
McNemar exact	0.0703	0.0525
McNemar mid-*p*	0.0391	0.0347
Exact unconditional	0.0353	0.0342

Third, the asymptotic McNemar test (without CC) performs surprisingly well, even for quite small sample sizes. It often violates the nominal significance level, but not by much. The largest type I error rate of the asymptotic McNemar test we observed in this study was 5.37% with a 5% nominal level. If that degree of infringement on the nominal level is acceptable, the asymptotic McNemar test is superior to the other tests. This is notably different from comparing two independent binomial proportions, where the asymptotic chi-squared test can produce substantial violations of the type I error rate in small samples [14].

The asymptotic test with CC performs similarly to—and sometimes even more conservatively than—the exact conditional test, and we do not recommend that it is used. This was expected, and is in line with the unequivocal recommendations against using the asymptotic chi-squared test with Yates's CC for the analysis of the independent 2×2 table [5,13,17].

We have only evaluated tests based on the McNemar statistic. It is also possible to construct tests using the

likelihood ratio statistic; however, Lloyd [18] found no practical difference between the two statistics. We prefer the much simpler—and widely used—McNemar statistic.

Conclusions

The McNemar mid-p test is a considerably improvement on the McNemar exact conditional test. The mid-p test did not violate the nominal level in any of the 9595 scenarios considered in this article and is thus an excellent alternative to the vastly more complex exact unconditional test. The most powerful test is the McNemar asymptotic test (without CC), which we recommend if small but frequent violations of the nominal level is acceptable. We do not recommend use of the McNemar exact conditional test nor the asymptotic test with CC in any situation.

Additional files

Additional file 1: How to calculate the McNemar mid-p test. This document shows how to calculate the McNemar mid-p test using the software packages Excel, Matlab, R, SAS, SPSS, Stata, StatsDirect, and StatXact.

Additional file 2: Box-plots of type I error rates from the evaluation study. This document shows box-plots of type I error rates from the total and various subregions of the evaluation study.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MWF conceived of the study, designed and carried out the evaluation of the tests, wrote an initial draft, and worked on the production of final draft. SL conceived of the study, participated in the design of the evaluation of the tests, and worked on the production of final draft. PL conceived of the study, participated in the design of the evaluation of the tests, and worked on the production of final draft. All authors read and approved the final manuscript.

Author details

[1]Unit of Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway. [2]Regional Centre for Child and Youth Mental Health and Child Welfare-Central Norway, Norwegian University of Science and Technology, Trondheim, Norway. [3]Department of Biostatistics, University of Oslo, Oslo, Norway.

Received: 12 March 2013 Accepted: 9 July 2013
Published: 13 July 2013

References

1. Bentur L, Lapidot M, Livnat G, Hakim F, Lidroneta-Katz C, Porat I, Vilozni D, Elhasid R: **Airway reactivity in children before and after stem cell transplantation.** *Pediatr Pulm* 2009, **44:**845–850.
2. Cavo M, Pantani L, Petrucci MT, Patriarca F, Zamagni E, Donnarumma D, Crippa C, Boccadoro M, Perrone G, Falcone A, Nozzoli C, Zambello R, Masini L, Furlan A, Brioli A, Derudas D, Ballanti S, Dessanti ML, De Stefano V, Carella AM, Marcatti M, Nozza A, Ferrara F, Callea V, Califano C, Pezzi A, Baraldi A, Grasso M, Musto P, Palumno A: **Bortezomib-Thalidomide-Dexamethasone is superior to Thalidomide-Dexamethasone as consolidation therapy after autologous hematopoietic stem cell transplantation in patients with newly diagnosed multiple myeloma.** *Blood* 2012, **120:**9–19.
3. Agresti A: **Dealing with discreteness: making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact.** *Stat Methods Med Res* 2003, **12:**3–21.
4. Lancaster HO: **Significance tests in discrete distributions.** *J Am Stat Assoc* 1961, **56:**223–234.
5. Lydersen S, Fagerland MW, Laake P: **Recommended tests for association in 2 x 2 tables.** *Stat Med* 2009, **28:**1159–1175.
6. Agresti A: *Categorical Data Analysis (3rd edn)*. Hoboken, NJ: Wiley, 2013.
7. McNemar Q: **Note on the sampling error of the difference between correlated proportions or percentages.** *Psychometrika* 1947, **12:**153–157.
8. Edwards AL: **Note on the "correction for continuity" in testing the significance of the difference between correlated proportions.** *Psychometrika* 1948, **13**(3):185–187.
9. Altman DG: *Practical Statistics for Medical Research*. Boca Raton Fl: Chapman & Hall/CRC, 1991.
10. Fleiss JL, Levin B, Paik MC: *Statistical Methods for Rates and Proportions (3rd edn)*. Hoboken, NJ: Wiley, 2003.
11. Mehta CR, Walsh SJ: **Comparison of exact, mid-p, and Mantel-Haenszel confidence intervals for the common odds ratio across several 2 x 2 contingency tables.** *Am Stat* 1992, **46:**146–150.
12. Lydersen S, Laake P: **Power comparison of two-sided exact tests for association in 2 x 2 contingency tables using standard, mid-p, and randomized test versions.** *Stat Med* 2003, **22:**3859–3871.
13. Hirji KF: *Exact Analysis of Discrete Data*. Boca Raton, FL: Chapman & Hall/CRC, 2006.
14. Hirji KF, Tan S-J, Elashoff RM: **A quasi-exact test for comparing two binomial proportions.** *Stat Med* 1991, **10:**1137—1153.
15. Suissa S, Shuster JJ: **The 2 x 2 matched-pairs trial: exact unconditional design and analysis.** *Biometrics* 1991, **47:**361–372.
16. Fagerland MW: **Exact and mid-p confidence intervals for the odds ratio.** *Stata J* 2012, **12:**505–514.
17. Haviland MG: **Yates's correction for continuity and the analysis of 2 x 2 contingency tables.** *Stat Med* 1990, **9:**363–367.
18. Lloyd CJ: **A new exact and more powerful unconditional test of no treatment effect from binary matched pairs.** *Biometrics* 2008, **64:**716–723.

doi:10.1186/1471-2288-13-91
Cite this article as: Fagerland *et al.*: The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology* 2013 **13**:91.