Journal of Clinical Epidemiology 68 (2015) 733–739

Journal of Clinical Epidemiology

The special case of the 2 × 2 table: asymptotic unconditional McNemar test can be used to estimate sample size even for analysis based on GEE

Cornelia M. Borkhoff[a,b,c,*], Patrick R. Johnston[d], Derek Stephens[e,f], Eshetu Atenafu[f,g]

[a]*Division of Pediatric Medicine and the Pediatric Outcomes Research Team (PORT), Department of Pediatrics and Child Health Evaluative Sciences, The Hospital for Sick Children, Peter Gilgan Centre for Research and Learning, 686 Bay St., Toronto, Ontario, M5G 0A4, Canada*

[b]*Women's College Research Institute, Women's College Hospital, 7th Floor, 790 Bay St., Toronto, Ontario, M5G 1N8, Canada*

[c]*Institute of Health Policy, Management and Evaluation, University of Toronto, 155 College St., Suite 425, Toronto, Ontario, M5T 3M6, Canada*

[d]*Clinical Research Program, Children's Hospital Boston, 300 Longwood Avenue, Boston, MA 02115, USA*

[e]*Child Health Evaluative Sciences, The Hospital for Sick Children, Peter Gilgan Centre for Research and Learning, 686 Bay St., Toronto, Ontario, M5G 0A4, Canada*

[f]*Department of Biostatistics, Dalla Lana School of Public Health, University of Toronto, 6th Floor, 155 College St., Toronto, Ontario, M5T 3M7, Canada*

[g]*Department of Biostatistics, Princess Margaret Cancer Center, University Health Network, 610 University Avenue, Toronto, Ontario, M5G 2M9, Canada*

Accepted 4 September 2014; Published online 13 November 2014

Abstract

Objectives: Aligning the method used to estimate sample size with the planned analytic method ensures the sample size needed to achieve the planned power. When using generalized estimating equations (GEE) to analyze a paired binary primary outcome with no covariates, many use an exact McNemar test to calculate sample size. We reviewed the approaches to sample size estimation for paired binary data and compared the sample size estimates on the same numerical examples.

Study Design and Setting: We used the hypothesized sample proportions for the 2 × 2 table to calculate the correlation between the marginal proportions to estimate sample size based on GEE. We solved the inside proportions based on the correlation and the marginal proportions to estimate sample size based on exact McNemar, asymptotic unconditional McNemar, and asymptotic conditional McNemar.

Results: The asymptotic unconditional McNemar test is a good approximation of GEE method by Pan. The exact McNemar is too conservative and yields unnecessarily large sample size estimates than all other methods.

Conclusion: In the special case of a 2 × 2 table, even when a GEE approach to binary logistic regression is the planned analytic method, the asymptotic unconditional McNemar test can be used to estimate sample size. We do not recommend using an exact McNemar test.

Keywords: Sample size; Two-period crossover trials; Subject-specific; Population-averaged; McNemar test; Generalized estimating equations

1. Introduction

When designing a study, an essential first step is to estimate the sample size required to have a reasonable power of detecting a hypothesized minimum clinically important difference, MCID, in the primary outcome variable, if it exists, at a given level of statistical significance. MCID is the smallest difference between experimental and control groups that one would consider to be clinically important. When sample size is too small, one risks failing to detect an MCID (type II error). When sample size is too large, finding a difference, which is not clinically important (but is statistically significant) (type I error), is more likely. Either case is an unethical use of subjects and waste of resources [1,2].

Paired binary data arise in studies with two measures on the same subject in before-and-after trials, two-period crossover trials, and matched case-control studies. For a binary outcome (eg, yes/no), a 2 × 2 table with the same row and column categories summarizes the data. Although repeated-measures designs can consist of multiple measures on the same subject, the focus of this article is on those

Conflict of interest: None.

Funding: This research was supported by grants from the Canadian Institute for Health Research (MOP-67765 and MOP-69081) and the Arthritis Society of Canada (99/093; renumbered to 99/0143 in 2001). C.M.B was supported by a Peterborough K.M. Hunter Graduate Studentship, a Canadian Arthritis Network Graduate Student Award and a Toronto Star Bursary Award.

* Corresponding author. Tel.: +1-416-813-7654x306978; fax: +1-416 813-5663.

E-mail address: cory.borkhoff@sickkids.ca (C.M. Borkhoff).

http://dx.doi.org/10.1016/j.jclinepi.2014.09.025
0895-4356/

What is new?

Key findings

- The asymptotic unconditional McNemar test can be used to estimate sample size for a paired binary outcome with no covariates even when the analysis is based on GEE.

What this study adds to what was known?

- The asymptotic unconditional McNemar test is a good approximation of Pan's GEE method. Sample size estimates based on exact McNemar are too conservative and yields unnecessarily larger sample size estimates than all other methods.

What is the implication? What should change now?

- The traditional advice is to use an exact McNemar test to estimate sample size for a paired binary outcome with no covariates. We recommend using the asymptotic unconditional McNemar test irrespective of the planned analysis.

designs with only two measures per subject, with no covariates. A key feature of paired binary data is that these two observations from the same individual tend to be positively correlated, which must be accounted for in the design and analysis of the study [3,4]. With the increasing use of marginal models, such as generalized estimating equation (GEE) models, and introduction of the procedure GLIMMIX for conditional models by SAS version 9.1, researchers may be unclear on how to estimate sample size for studies with a paired binary primary outcome.

To not commit, what is an example of a type III error (giving the right answer to the wrong question) [5], the approach to sample size estimation should align with the planned statistical analysis of the primary outcome [6,7]. Approaches to analyzing paired binary data can be grouped into two classes of logistic regression models: subject-specific (or conditional) models and population-averaged (or marginal) models [3,4,8−10]. A sample size estimate based on a McNemar test would be the method of choice when estimating subject-specific effects [11]. But which McNemar test? A sample size estimate based on a GEE method is used when estimating population-averaged effects [12−15]. Identifying the nature of your research question at the design stage will define the method of analysis, and hence, the formula one should use to estimate sample size.

The purpose of this article was (1) to review the two major approaches to sample size estimation for paired binary data and describe how they differ; and 2) to compare the sample size estimates based on a McNemar test vs. GEE on the same numerical examples. In this article, we suggest that in the special case of a 2×2 table, the sample size estimation method need not align with the planned analysis. An asymptotic unconditional McNemar test can be used to estimate sample size irrespective of the planned analysis.

2. A motivating example

The motivation for this comparison of approaches to sample size estimation for paired binary data came from a study examining how patient gender affected physicians' treatment recommendations regarding total knee arthroplasty (TKA) [16]. Specifically, two standardized patients (one man and one woman) differing only in gender with otherwise identical case histories underwent blinded assessments by family physicians and orthopedic surgeons located in Ontario, Canada. Both standardized patients presented with chronic knee pain as their chief complaint and level of function, pain, and prior treatment appropriate for a patient with moderate knee osteoarthritis. At the end of the visit, standardized patients recorded the binary primary outcome: the physicians' recommendations for TKA (yes = 1 and no = 0) on a postvisit checklist.

Sample size was based on a one-sided alternative hypothesis that physicians would be less likely to recommend TKA to the female patient compared with the male patient. By convention, the type I error rate or α was set at 0.05 and the type II error rate or β was selected to be 0.20. We estimated the magnitude of the hypothesized effect of patients' gender on the physicians' treatment recommendations based on findings from a population-based study, which showed a greater than threefold gender disparity in access to total joint arthroplasty (5.3 per 1,000 vs. 1.6 per 1,000 for women and men, respectively) [17]. To obtain an odds ratio of 3.3, we assumed that 30.7% of physicians would recommend TKA to the man but not the woman and that 9.3% would recommend the procedure to the woman but not the man. For the remaining physicians, we assumed that 30% would recommend TKA to both the man and the woman and that 30% would recommend neither patient for TKA. We based our sample size on the exact McNemar test for paired proportions; each pair consisted of the female and the male patient visiting each physician. We determined that a sample size, η, of 58 physicians was required and assuming a 15% dropout rate, 71 physicians needed to be enrolled ($58/(1 - 0.15) = 71$).

Sample size estimation based on GEE requires several parameters, all of which were unknown at the design stage of the study. These included the correlation of physicians' recommendations for TKA, specifying the structure of the correlation matrix, and a scale factor allowing for extra variation (or dispersion) in the response beyond the assumed variance [12]. For these reasons, although we planned to analyze our data using GEE, we based our sample size of $\eta = 58$ on what was considered to be a

conservative estimate using the exact McNemar test. Did we commit a type III error?

We based our analytic approach on the nature of our research question. We were interested in answering the ''population-averaged'' question that is—What is the probability that physicians would recommend TKA to a man over a woman? This approach is used when the results are important from a population health perspective. The corresponding ''subject-specific'' question would be—What is the probability that a given physician would recommend TKA to a man vs. a woman? This approach would be used to help a clinician decide whether to recommend treatment to their patient, for example.

Despite our study data being reduced to a 2×2 table, a GEE approach to binary logistic regression was used for statistical analysis. Was it appropriate to use the exact McNemar test to estimate sample size when we planned to analyze our data using the GEE approach? Is the sample size estimate based on an exact McNemar test even conservative? What would the sample size estimate based on GEE have been for our study?

3. Approaches to sample size estimation for paired binary data

Approaches to sample size estimation for paired binary data can be grouped into two classes of logistic regression models: subject-specific and population-averaged models [3,4,8–10]. These two analytic approaches differ in how they account for the correlated repeated measures on the same subject (or cluster of subjects sharing the same treatment effect) and in how regression coefficients are estimated and interpreted.

3.1. The subject-specific approach

Conditional logistic (fixed-effects) regression models are commonly used in a subject-specific approach to correlated data analysis [4,18]. Conditional maximum likelihood estimation is used with the likelihood conditioned on the discordant pairs (the observations where [$y_{i1} = 1$ and $y_{i2} = 0$] and [$y_{i1} = 0$ and $y_{i2} = 1$]) [10,18]. Although the large number of intercepts (α_i) considered nuisance parameters are conditioned out, subject-to-subject heterogeneity is still explicitly accounted for [8]. For our study, this model to describe the probability of an individual physician recommending TKA can be written as:

$$\text{logit}(E[Y_{it}|\alpha_i]) = \alpha_i + \beta_C \text{GENDER}_{it}$$

where Y_{it} denotes a binary outcome (recommended TKA, yes = 1 and no = 0) for physician i at time t; $E[Y_{it}|\alpha_i]$ denotes the expectation or the mean of the response for a given physician; GENDER_{it} denotes the gender of the standardized patient (male = 1 and female = 0) for physician i at time t. In conditional models, the regression coefficients, β_C, are conditional on the physician and describe the average response to changing patient gender for an individual physician or cluster of physicians. The effect is subject-specific because it is defined at the subject level or in our case, at the level of the physician. Table 1 is the subject-specific display of paired data, whereby the responses of each physician participating in our study can be described by one of four possible partial tables.

In subject-specific models, subjects sharing the same partial table and thereby the same treatment effect have their own probability distributions, which are defined by their same intercept (α_i) in the model [18]. With the intercept (α_i) as a fixed constant over repeated observations, the correlation for the paired binary data within subjects is explicitly taken into account. Paired observations for the same physician have the same intercept; thus, a physician who tends to recommend TKA to a man tends to also recommend TKA to a woman.

A second conditional approach to correlated data analysis is random-effects models (also known as hierarchical linear models). Often called random-intercept models, rather than treating the intercepts (α_i) as fixed constants, they are treated as random variables with a specified distribution (usually normal) [3,19].

Conditional logistic regression reduces to the test by McNemar when a single binary explanatory variable is used (with no covariates) [20]. A sample size estimate based on a test by McNemar would be the method of choice when estimating subject-specific effects. If $\{p_{11}, p_{12}, p_{21}, p_{22}\}$ represent the sample proportions for a 2×2 table, then the test of $H_O : p_{12} = p_{21}$ is the chi-square test by McNemar statistic with 1° of freedom:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

To estimate the required sample size based on the test by McNemar, one needs to either specify p_{12} and p_{21} or equivalently the sum of discordant pairs $p_{12} + p_{21}$ and the difference in discordant pairs $p_{12} - p_{21}$ or the odds ratio p_{12}/p_{21}.

Table 1. Subject-specific display of paired data

	Recommend TKA		
Gender	**No**	**Yes**	**Total**
(No, no) configuration table			
Man	1	0	1
Woman	1	0	1
(No, yes) configuration table			
Man	1	0	1
Woman	0	1	1
(Yes, no) configuration table			
Man	0	1	1
Woman	1	0	1
(Yes, yes) configuration table			
Man	0	1	1
Woman	0	1	1

Abbreviation: TKA, total knee arthroplasty.

There are a number of different sample size formulas based on the test by McNemar. These can be classified as either exact (binomial probability distribution) or asymptotic (normal approximation to the binomial distribution) and conditional or unconditional [21,22]. The conditional procedure is based only on the proportion of discordant pairs, whereas the unconditional procedure uses all the matched pairs, including the concordant pairs [21].

In this study, we have chosen three formulas with which to compare the GEE approach to sample size estimation for paired binary data. The first is the exact McNemar (also known as the conditional binomial test) [23], used by the PASS software (Kaysville, UT, USA) [24] and the one we used to estimate sample size. The other two formulas are the unconditional version of the asymptotic test [25] and the conditional version of the asymptotic test [26], see Appendix A at www.jclinepi.com for sample size formulas. The unconditional version is known to be somewhat more conservative than the conditional version [21]. Lachin [27] wrote a SAS macro to calculate sample size based on the asymptotic unconditional McNemar test (available at http://www2.bsc.gwu.edu/bsc/docs/JohnLachin/mcnemarn.sas).

3.2. The population-averaged approach

In the population-averaged approach, the treatment effects are estimated as a function of covariates without explicitly accounting for subject-to-subject heterogeneity [8]. The GEE method to binary logistic regression is the most common approach to estimation. GEEs were proposed as an extension to generalized linear models to perform regression analyses on correlated data arising from distributions that are not normally distributed (such as a binomial distribution) [28]. This model can be written as:

$$\text{logit}(E[Y_{it}]) = \alpha + \beta_M \text{GENDER}_{it}$$

where $E[Y_{it}]$ denotes the marginal expectation or the marginal means of the response. A marginal treatment effect is the average treatment effect on the population. As the response is binary, the GEE method uses a logit transformation of the marginal expectation of response to create a linear function between the log odds of the response and the explanatory variable(s), in this case patient gender. In marginal models, the regression coefficients, β_M, describe the average response over the subpopulation that shares a common value of patient gender [29]. Thus, β_M is completely different from β_C in the conditional model. In this model, the exponentiated value of β_M (written as $\exp(\beta_M)$) is the odds ratio of the event (recommending TKA) among physicians when GENDER = 1 compared with when GENDER = 0. The treatment effects in these marginal models are population-averaged and cannot provide estimates of changes within individuals over time. The marginal totals $\Sigma_i y_{i1} = (a+b)$ and $\Sigma_i y_{i2} = (a+c)$ are referred to as marginal proportions, p_1 and p_2 and in this case are the average recommendation rates for a man and woman, respectively. The yes, no and no, yes configuration tables in Table 1 represent the discordant pairs (p_{12} and p_{21}) in the population-averaged display of paired data in Table 2, respectively. The marginal proportions in Table 2 are the sum of the elements in the 2 × 2 table that results from collapsing the four partial tables in Table 1 [30].

Note that the intercept (α) in the model is no longer subject-specific. To account for the correlation between the paired observations, a working correlation matrix for the vector of repeated observations from each subject (assumed to be the same for all subjects) is specified. The working correlation matrix is analogous to the intercepts (α_i) in the subject-specific model that account for the correlated paired binary data.

There are primarily two sample size formulas based on GEE: one based on a score χ^2 test statistic derived by Liu and Liang [13] and one based on a Wald χ^2 test statistic derived explicitly for two-period crossover trials by Pan [14] extending the previous work by Shih [15], whose formula was for the parallel group design (see 4 and 5 in Appendix A at www.jclinepi.com). Rochon [12] wrote a SAS macro to calculate sample size based on a Wald χ^2 test statistic for parallel group designs. Dahmen and Ziegler [31] extended this SAS macro to two-period crossover trials (available at http://www.imbs-luebeck.de/imbs/de/node/30).

4. Illustration comparing approaches to sample size estimation for paired binary data

We estimated sample size using an exact McNemar test as it was considered conservative, and sample size estimation based on GEE involves assigning values to three design parameters—all of them unknown at the start of our study. As it happens, one can use a conservative approach by setting the correlation equal to zero [as the correlation decreases, the variance increases (variance = σ^2 (1-ρ)) and so does the sample size] and setting the scale factor equal to 1 [12,29,32]. With exactly two observations per subject, all correlation structures

Table 2. Population-averaged display of paired data

Recommendation regarding total knee arthroplasty to male standardized patient	Recommendation regarding total knee arthroplasty to female standardized patient		
	Yes	No	Total
Yes	a (p_{11})	b (p_{12})	a + b (p_1)
No	c (p_{21})	d (p_{22})	c + d ($1-p_1$)
Total	a + c (p_2)	b + d ($1-p_2$)	n

would yield the same working correlation matrix. Sample size estimation involves estimating the hypothesized MCID; doing so, usually results in specifying hypothesized sample proportions $\{p_{11}, p_{12}, p_{21}, p_{22}\}$ for the 2 × 2 table. Rather than set the correlation to zero, one can use the hypothesized sample proportions to calculate the correlation between the marginal proportions using the formula [33] shown in Appendix B at www.jclinepi.com. The correlation for our standardized patient study is equal to 0.25759.

In Table 3, the required sample sizes are given for the three different McNemar tests and for the two sample size formulas based on GEE. When the correlation is equal to 0.25759, the asymptotic unconditional McNemar test [25] leads to identical sample sizes as GEE method by Pan [14] based on a Wald χ^2 test statistic, suggesting that the asymptotic unconditional McNemar test is also a Wald χ^2 test statistic. The asymptotic conditional McNemar test [26] leads to identical sample sizes as GEE method by Liu and Liang [13] based on a score χ^2 test statistic, suggesting that the asymptotic conditional McNemar test is also a score χ^2 test statistic. Sample size estimates based on a Wald χ^2 test statistic are higher than those based on a score χ^2 test statistic [21,31]. The exact McNemar yields the highest sample size estimates. As expected, when using GEE methods, setting the correlation close to zero yields higher sample sizes than when correlation was equal to 0.25759. However, setting the correlation to zero is inappropriate as doing so completely changes the hypothesized sample proportions, the sum, and the difference of the discordant pairs and the odds ratio.

To understand to what extent our findings generalize, we looked to another numerical example. Table 4 is an extension of a table presented by Dahmen and Ziegler [31], comparing GEE method by Pan [14] with GEE method by Liu and Liang [13]. We extended this table to compare sample size estimates for a two-period crossover study with varying r and marginal proportions.

Using a SAS program written by Patrick Johnston (see Appendix C at www.jclinepi.com), we solved what the hypothesized sample proportions for the 2 × 2 table would be, based on the correlation r and the marginal proportions (p_1 and p_2), then calculated the sample size estimates for the three different subject-specific approaches. Required sample size decreases when the correlation between marginal proportions increases, as previously shown in Table 3. In this example, the asymptotic unconditional McNemar test [25] leads to almost identical sample sizes as GEE method by Pan [14]. In only 2 of 12 instances in Table 4, the sample size estimate based on the asymptotic unconditional McNemar test was lower by more than $n = 1$, suggesting that the asymptotic unconditional McNemar test is a good approximation of GEE method by Pan. The settings in which these methods do not yield similar results are when both marginal proportions are less than or equal to 0.2 and when there is a low correlation between the marginal proportions [eg, for $r = 0.1$, $p_1 = 0.2$, $p_2 = 0.1$, the sample size estimate based on the asymptotic unconditional McNemar test is lower than the sample size estimate based on GEE method by Pan ($n = 5$)]. We did not find sample size estimates based on asymptotic conditional McNemar test [26] to be a good approximation of those based on GEE method by Liu and Liang [13]. Again, using an exact McNemar leads to considerably higher sample size estimates than all other methods.

5. Discussion

Our article is a review of the two major approaches (subject-specific vs. population-averaged) for calculating the required sample size for a paired binary outcome with no covariates. Aligning the method used to estimate sample size with the planned analytic method ensures that actual power matches the planned power. We contribute an

Table 3. Sample size estimates for standardized patient study based on the subject-specific and population-averaged approaches for paired binary data for 80% power at a significance level of 5%

	Sample size estimates						
	Subject-specific			Population-averaged			
Model parameters	Exact "conditional" McNemar	Asymptotic unconditional McNemar	Asymptotic conditional McNemar	Pan GEE	Liu and Liang GEE	Pan GEE	Liu and Liang GEE
One sided							
η	58	52	48	52	48	70	64
Two sided							
η	73	66	62	66	61	88	82
Known parameters							
P_1		0.607		0.607	0.607	0.607	0.607
P_2		0.393		0.393	0.393	0.393	0.393
r	0.25759	0.25759	0.25759	0.25759	0.25759	0	0
P_{11}		0.3					
P_{12}	0.307	0.307	0.307				
P_{21}	0.093	0.093	0.093				
P_{22}		0.3					
		Wald χ^2	Score χ^2	Wald χ^2	Score χ^2	Wald χ^2	Score χ^2

Abbreviation: GEE, generalized estimation equation.

Table 4. Sample size estimates for a two-period crossover study design based on the subject-specific and population-averaged approaches for paired binary data for 80% power at a significance level of 5%

			Sample size estimates				
			Subject-specific			Population-averaged	
Model parameters			Exact "conditional" McNemar	Asymptotic unconditional McNemar	Asymptotic conditional McNemar	Pan GEE	Liu and Liang GEE
r=0.1	P_1=0.2	P_2=0.1	195	183	175	188	178
	P_1=0.3	P_2=0.2	283	268	263	269	262
	P_1=0.4	P_2=0.3	340	324	320	324	318
	P_1=0.5	P_2=0.4	370	352	349	351	347
r=0.3	P_1=0.2	P_2=0.1	159	146	135	148	140
	P_1=0.3	P_2=0.2	223	210	203	210	205
	P_1=0.4	P_2=0.3	269	253	248	252	248
	P_1=0.5	P_2=0.4	291	275	271	273	270
r=0.5	P_1=0.2	P_2=0.1	114	108	91	108	103
	P_1=0.3	P_2=0.2	166	152	142	151	147
	P_1=0.4	P_2=0.3	195	183	175	180	177
	P_1=0.5	P_2=0.4	211	198	191	195	193
				Wald χ^2	Score χ^2	Wald χ^2	Score χ^2

Abbreviation: GEE, generalized estimation equation.
Extension of Table 4 from Dahmen and Ziegler [31].
In all examples, the difference in marginal proportions = $P_1 - P_2$ is 0.1.

important and practical finding that in the special case of a 2 × 2 table, the sample size estimation method need not align with the planned analysis. An asymptotic unconditional McNemar test can be used to estimate sample size even when GEE is the planned analytic method, as it is a good approximation. The traditional advice is to use an exact McNemar test. However, we also observed that the exact McNemar is too conservative (yields unnecessarily large sample sizes) and recommend that it not be used as SAS programs to estimate sample size using the asymptotic unconditional McNemar test and GEE method by Pan are readily available.

Our study's sample size of 58 physicians was determined using an exact McNemar test. The best would have been to base sample size on GEE method by Pan [14] to align with our planned analysis. We recommend a two-step process when estimating sample size for paired binary data when GEE is the planned analytic method. First, one needs to specify the marginal proportions and rather than specify the correlation as zero, calculate the correlation between the marginal proportions using the SAS program in Appendix C at www.jclinepi.com. And second, use the SAS macro for a two-period crossover design available http://www.imbs-luebeck.de/imbs/de/node/30 [31]. The required sample size based on GEE method by Pan for our study was 52. As sample size estimates based on the asymptotic unconditional McNemar test are good approximations of those based on GEE method by Pan, this test is a simpler alternative. This test is both convenient and practical, as one need only specify the inside proportions. A SAS macro to calculate sample size based on the asymptotic unconditional McNemar test available at http://www2.bsc.gwu.edu/bsc/docs/JohnLachin/mcnemarn.sas [27]. The required sample size based on the asymptotic unconditional McNemar test for our study was 52 (odds ratio = 3.3, $p_{21} = 0.093$, $\alpha = 0.05$, and $\beta = 0.20$).

For our study, we originally proposed to send two pairs of standardized patients to each physician (with the second pair being one man and one woman with severe osteoarthritis). All the sample size formulas mentioned in this article are for one level of clustering. Approaches do exist for more than one level of clustering [34,35]; however, further research is required to extend these sample size estimation methods to crossover trials.

Perhaps the best approach is to use Monte Carlo or bootstrap simulation, as one can estimate sample size based on almost any statistical test [36–38]. Some disadvantages are that simulation requires computational expertise and can be time consuming. The advantage is that one can include all covariates in the model and with such an explicit sample size formula, calculate the most efficient sample size.

6. Conclusion

The approach to sample size estimation for paired binary data depends on the nature of your research question. In turn, this will determine the approach to statistical analyses. For analyses based on a subject-specific approach, we recommend estimating sample size based on the asymptotic unconditional McNemar test. Even for analyses based on a GEE approach to binary logistic regression, in the special case of a 2 × 2 table, the asymptotic unconditional McNemar test can be used to estimate sample size. We do not recommend using an exact McNemar test.

Acknowledgments

The authors would also like to thank George Tomlinson for his helpful comments on an earlier version of this manuscript.

Authors' contributions: All authors contributed to the sample size estimation of the numerical examples described in the article and the interpretation of the different approaches to sample size. C.M.B. drafted the article and all authors participated in its critical revision for important intellectual content and gave final approval of the submitted manuscript.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2014.09.025.

References

[1] Altman DG. Statistics and ethics in medical research III: how large a sample? BMJ 1980;281:1336−8.
[2] Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288:358−67.
[3] Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. Am J Epidemiol 1998;147:694−703.
[4] Allison P. Logistic regression using the SAS system: theory and application. Cary, N.C.: SAS Institute Inc. 1999.
[5] Kimball AW. Errors of the third kind in statistical consulting. J Am Stat Assoc 1957;52:133−42.
[6] Delucchi KL. Sample size estimation in research with dependent measures and dichotomous outcomes. Am J Public Health 2004;94:372−7.
[7] Dang Q, Mazumdar S, Houck PR. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. Comput Methods Programs Biomed 2008;91(2):122−7.
[8] Zeger SL, Liang KY, Albert PS. Methods for longitudinal data: a generalized estimating equation approach. Biometrics 1988;44:1049−60.
[9] Zorn CJW. Generalized estimating equation models for correlated data: a review with applications. Am J Polit Sci 2001;45(2):470−90.
[10] Stokes ME, Davis CS, Koch GG. Categorical data analysis using the SAS system. Cary, N.C.: SAS Institute Inc. 2000.
[11] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153−7.
[12] Rochon J. Application of GEE procedures for sample size calculations in repeated measures experiments. Stat Med 1998;17:1643−58.
[13] Liu G, Liang KY. Sample size calculations for studies with correlated observations. Biometrics 1997;53:937−47.
[14] Pan W. Sample size and power calculations with correlated binary data. Control Clin Trials 2001;22:211−27.
[15] Shih WJ. Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimating equations. Biometrical J 1997;399(8):899−908.
[16] Borkhoff CM, Hawker GA, Kreder HJ, Glazier RH, Mohamed NN, Wright JG. The effect of patients' sex on physicians' recommendations for total knee arthroplasty. CMAJ 2008;178(6):681−7.
[17] Hawker GA, Wright JG, Coyte PC, Williams JI, Harvey B, Glazier R, et al. Differences between men and women in the rate of use of hip and knee arthroplasty. N Engl J Med 2000;342:1016−22.
[18] Agresti A. An introduction to categorical data analysis. New York, NY: John Wiley & Sons, Inc. 1996.
[19] Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Repeated measures and longitudinal data analysis (chapter 8). Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. New York, NY: Springer; 2005:209−40.
[20] Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45:562−5.
[21] Sahai H, Khurshid A. Formulas and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the matched pair design: a review. Fundam Clin Pharmacol 1996;10(6):554−63.
[22] Royston P. Exact conditional and unconditional sample size for pair-matched studies with binary outcome: a practical guide. Stat Med 1993;12:699−712.
[23] Schork MA, Williams GW. Number of observations required for the comparison of two correlated proportions. Comm Stat Simulat Comput 1980;B9:349−57.
[24] Hintze JL. PASS 2005: power analysis and sample size. Kaysville, UT: NCSS; 2005.
[25] Connett JE, Smith JA, McHugh RB. Sample size and power for pair-matched case-control studies. Stat Med 1987;6:53−9.
[26] Schlesselman JJ. Case-control studies. New York, NY: Oxford University Press; 1982.
[27] Lachin JM. Biostatistical methods: the assessment of relative risks. New York, NY: John Wiley & Sons, Inc. 2000.
[28] Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73(1):13−22.
[29] Diggle PJ, Heagerty P, Liang KY, Zeger SL. Analysis of longitudinal data. Oxford, U.K.: Oxford University Press; 2002.
[30] Agresti A, Min Y. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. Stat Med 2004;23:65−75.
[31] Dahmen G, Ziegler A. Generalized estimating equations in controlled clinical trials: hypothesis testing. Biometrical J 2004;46(2):214−32.
[32] Connor RJ. Sample size for testing differences in proportions for the paired-sample design. Biometrics 1987;43:207−11.
[33] Lehr RG. Some practical considerations and a crude formula for estimating sample size for McNemar's test. Drug Inf J 2001;35(4):1227−33.
[34] Eliasziw M, Donner A. Application of the McNemar test to non-independent matched pair data. Stat Med 1991;10:1981−91.
[35] Liu A, Shih WJ, Gehan E. Sample size and power determination for clustered repeated measurements. Stat Med 2002;21:1787−801.
[36] Muthen LK, Muthen BO. How to use a Monte Carlo study to decide on sample size and determine power. Struct Equ Modeling 2002;9(4):599−620.
[37] Walters SJ, Campbell MJ. The use of bootstrap methods for estimating sample size and analyzing health-related quality of life outcomes. Stat Med 2005;24:1075−102.
[38] Feiveson AH. Power by simulation. Stata J 2002;2:107−24.