Nonparametric Sample Size Estimation for Sensitivity and Specificity with Multiple Observations per Subject

Fan Hu

Department of Statistical Science

Southern Methodist University

Dallas, TX

William R. Schucany

Department of Statistical Science

Southern Methodist University

Dallas, TX

Chul Ahn

Department of Clinical Sciences

UT Southwestern Medical Center

Dallas, TX

Correspondence should be sent to:

Chul Ahn, Ph.D.

Department of Clinical Sciences

UT Southwestern Medical Center

5323 Harry Hines Blvd, E5.506

Dallas, TX 75390

email: Chul.Ahn@UTSouthwestern.edu

Nonparametric Sample Size Estimation for Sensitivity and Specificity with Multiple Observations per Subject

Summary

We propose a sample size calculation approach for the estimation of sensitivity and specificity of diagnostic tests with multiple observations per subjects. Many diagnostic tests such as diagnostic imaging or periodontal tests are characterized by the presence of multiple observations for each subject. The number of observations frequently varies among subjects in diagnostic imaging experiments or periodontal studies. Nonparametric statistical methods for the analysis of clustered binary data have been recently developed by various authors. In this paper, we derive a sample size formula for sensitivity and specificity of diagnostic tests using the sign test while accounting for multiple observations per subjects. Application of the sample size formula for the design of a diagnostic test is discussed. Since the sample size formula is based on large sample theory, simulation studies are conducted to evaluate the finite sample performance of the proposed method. We compare the performance of the proposed sample size formula with that of the parametric sample size formula that assigns equal weight to each observation. Simulation studies show that the proposed sample size formula generally yields empirical powers closer to the nominal level than the parametric method. Simulation studies also show that the number of subjects required increases as the variability in the number of observations per subject increases and the intracluster correlation increases.

Keywords: Intracluster correlation; Diagnostic test; Empirical power.

1 Introduction

Diagnostic tests are of particular importance in medicine since early and accurate diagnosis can decrease morbidity and mortality rates of disease. Some examples of diagnostic tests include positron emission tomography (PET) scans, X-rays, and enzymatic diagnostic tests. The performance of a diagnostic test is often summarized by its sensitivity and specificity. Sensitivity is defined as the probability of a positive diagnostic test in a subject with the disease, and specificity as the probability of a negative diagnostic test in a subject without the disease. In this paper we focus on clustered binary observations, which are made from multiple observations on each subject. In this case, observations from each subject are correlated although those from different subjects are independent. For example in a radiologic study each subject may contribute multiple lesions to the study and an observation is made from each lesion.

Fleiss *et al.* (1) discussed the estimation of sensitivity and specificity for independent observations. Ignoring the within-subject correlation may result in underestimation of the variance of sensitivity and specificity estimates. Hujoel *et al.* (2) emphasized the importance of incorporating correlation among observations citing the high priority of diagnostic tests by the National Institute of Dental Research. Hujoel *et al.* (2) proposed a correlated binomial model to estimate the sensitivity and specificity of the diagnostic test with multiple observations per subject. Smith and Hadgu (3) proposed a generalized estimating equation (GEE) approach to estimate the sensitivity and specificity in the presence of multiple observations per subject. Ahn (4, 5) evaluated various statistical methods for the estimation of sensitivity and specificity of diagnostic tests through simulation, and recommended statistical methods based on the values of intracluster correlation coefficients. Jung and Ahn (6) proposed an optimal weight estimator for the estimation of sensitivity and specificity, which minimizes the variance of the estimator.

In this paper we focus on sample size estimation for testing a binomial proportion producing a desired sensitivity or specificity of a diagnostic test with multiple observations per subject. Here, the sample size refers to the number of subjects. For the estimation of sample size under clustered binary data, we encounter two problems that need to be accommodated. One problem is that the number of observations within a subject may vary among subjects. The other problem is assessing the correlation among observations with a subject. Various authors provide sample size formulae for the design of clustered binary data. Parametric sample size formulae have been proposed by

many authors (7-12). Semiparametric methods, such as generalized estimating equation (GEE), have been used for the derivation of a sample size formula. Liu and Liang (13) derived the sample size formula using the score test from GEE whereas Pan (14) derived the formula using the $z-$test, a special case of the Wald test.

Nonparametric statistical methods for the analysis of clustered data have been recently developed by various authors (15-20). However, to our knowledge, there is no publication on sample size estimation with nonparametric approaches for testing a binomial proportion with multiple observations per subject. Noether (21) discussed sample size determinations for some common traditional nonparametric tests assuming that the observations are mutually independent. Here, we propose a sample size calculation method for clustered binary data using a sign test, which incorporates the intracluster correlation coefficient and variability in the number of observations per subject. Noether's sample size formula for a sign test is a special case of the sample size formula presented in this paper. We apply the proposed sample size formula to the design of a dental study. Extensive simulation studies are conducted to evaluate the performance of the sample size formula and to investigate the effects of intracluster correlation and imbalance in the number of observations per subject.

2 Statistical method

Let n be the total number of subjects in an experiment and m_i be the number of observations in the ith ($i = 1, ..., n$) subject. The number of observations per subject may vary at random with a certain distribution. Let X_{ij} be the binary random variable of the jth observation in the ith subject, $j = 1, ..., m_i$, which is coded as 1 for success and -1 for failure. For the estimation of sensitivity, success is defined as the positive diagnostic test in a subject with the disease, and failure as the negative diagnostic test in a subject with the disease. Success and failure are defined similarly for the estimation of specificity. We use this coding scheme since we can obtain some desirable properties. With this coding we can express the total as the difference between the total number of successes and the total number of failures for each subject (22, 23). We assume that observations within a subject are exchangeable in the sense that, given m_i, $X_{i1}, ..., X_{im_i}$ have a common marginal response probability $P(X_{ij} = 1) = p(0 < p < 1)$ and a common pairwise intracluster correlation

coefficient $\rho = \text{corr}(X_{ij}, X_{ij'})$ for $j \neq j'$. This correlation is assumed not to vary with the number of observations per subject. We test the null hypothesis $H_0 : p = p_0$ versus $H_1 : p = p_1$ for $p_0 \neq p_1$.

Let m_i^+ and m_i^- denote the total number of successes and failures in the ith subject, respectively. In this paper we consider the choice of a sequence of weights such that all observations receive the same weight. The statistic we use in this sign test is

$$T = \frac{n \sum_{i=1}^n S_i}{\sum_{i=1}^n m_i} = \frac{n \sum_{i=1}^n \left(m_i^+ - m_i^-\right)}{\sum_{i=1}^n m_i}, \tag{1}$$

where $S_i = \sum_{j=1}^{m_i} X_{ij}$.

The expected value of T under the null hypothesis is given by

$$\mu_0(T) = \frac{n \sum_{i=1}^n E(S_i|H_0)}{\sum_{i=1}^n m_i} = n(2p_0 - 1). \tag{2}$$

The variance of T under the null hypothesis is

$$\sigma_0(T)^2 = \frac{n^2 \sum_{i=1}^n Var(S_i|H_0)}{(\sum_{i=1}^n m_i)^2} = 4p_0(1-p_0)n^2 \frac{\sum_{i=1}^n m_i\{1 + (m_i - 1)\rho\}}{(\sum_{i=1}^n m_i)^2}, \tag{3}$$

which can be consistently estimated by

$$\widehat{\sigma_0(T)}^2 = 4p_0(1-p_0)n^2 \frac{\sum_{i=1}^n m_i\{1 + (m_i - 1)\hat{\rho}\}}{(\sum_{i=1}^n m_i)^2}, \tag{4}$$

where $\hat{\rho}$ can be obtained by the ANOVA method (9,24). The ANOVA method suitable for continuous variables can be adapted to estimate intracluster correlation coefficient for binary outcomes. The intracluster correlation coefficient is estimated by $(MSB - MSW)/[MSB + (\bar{m} - 1)MSW]$, where $\bar{m}$ is the average number of observations per subject, and MSB and MSW are the mean squares between and within clusters, respectively. Ridout *et al.* (25) conducted simulation studies to evaluate the performance of various estimators of ρ for clustered binary data under the common-correlation model, $\rho = \text{corr}(X_{ij}, X_{ij'})$ for $j \neq j'$. They showed that the ANOVA estimator performed well in their simulation studies.

The test statistic

$$Z = \frac{\sum_{i=1}^n \left(m_i^+ - m_i^-\right) - \sum_{i=1}^n m_i(2p_0 - 1)}{\sqrt{4p_0(1-p_0) \sum_{i=1}^n m_i\{1 + (m_i - 1)\hat{\rho}\}}} \tag{5}$$

is asymptotically normal with mean 0 and variance 1. Hence, we reject H_0 if the absolute value of Z is larger than $z_{1-\alpha/2}$, which is the $100(1 - \alpha/2)$ percentile of the standard normal distribution.

3 Sample size calculation with equal numbers of observations per subject

Noether (21) proposed a sample size determination for some common nonparametric tests such as a sign test. He showed that the sample size or the power of the test can be estimated by solving the following equation.

$$\left\{\frac{\mu_1(T)-\mu_0(T)}{\sigma_0(T)}\right\}^2 = (z_{1-\alpha/2}+rz_{1-\beta})^2,$$

where $r=\sigma_1(T)/\sigma_0(T)$.

When all subjects contribute equal number of observations, we have $m_i = m$ for $i = 1, ..., n$. The expectation of T under the null hypothesis is

$$\mu_0(T) = \sum_{i=1}^{n}\frac{1}{m}E(S_i|H_0) = n(2p_0-1).$$

Simarly, the expectation of T under the alternative hypothesis is

$$\mu_1(T) = \sum_{i=1}^{n}\frac{1}{m}E(S_i|H_1) = n(2p_1-1).$$

The variance of T under the null hypothesis is

$$\sigma_0(T)^2 = \sum_{i=1}^{n}\frac{1}{m^2}Var(S_i|H_0) = 4p_0(1-p_0)n\{1+(m-1)\rho\}/m.$$

To achieve a power of $1-\beta$, the required sample size (n) to test $H_0 : p = p_0$ against $H_1 : p = p_1$ is given by

$$n = \frac{(z_{1-\alpha/2}+rz_{1-\beta})^2}{(p_1-p_0)^2}\left\{\frac{1+(m-1)\rho}{m}\right\}p_0(1-p_0), \tag{6}$$

where $r=\sigma_1(T)/\sigma_0(T) = \sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}$. Throughout this paper, the sample size refers to the number of clusters.

4 Sample size determination with varying numbers of observations per subject

We assume that the number of observations (m_i) is small relative to the number of subjects (n), so that asymptotic results can be obtained with respect to n. The expectation of T under the null

and alternative hypotheses are

$$\mu_0(T) = \frac{n\sum_{i=1}^n E(S_i|H_0)}{\sum_{i=1}^n m_i} = n(2p_0 - 1)$$

and

$$\mu_1(T) = \frac{n\sum_{i=1}^n E(S_i|H_1)}{\sum_{i=1}^n m_i} = n(2p_1 - 1),$$

respectively.

The variance of T under the null distribution is

$$\sigma_0(T)^2 = \frac{n^2\sum_{i=1}^n Var(S_i|H_0)}{(\sum_{i=1}^n m_i)^2} = 4p_0(1-p_0)n^2\frac{\sum_{i=1}^n m_i\{1+(m_i-1)\rho\}}{(\sum_{i=1}^n m_i)^2}.$$

One can model the m_i's as independent and identically distributed random variables. From Equation (3)

$$\frac{1}{n}\sigma_0(T)^2 \to 4p_0(1-p_0)\{(1-\rho)E(M) + E(M^2)\rho\}/E(M)^2$$

as $n \to \infty$, where M is the random variable associated with the number of observations per subject and $E(.)$ is the expectation with respect to the distribution of the number of observations per subject. With $E(M) = \theta$, $Var(M) = \tau^2$, and $\gamma = \tau/\theta$, $\sigma_0(T)^2$ converges to

$$4np_0(1-p_0)\left\{\frac{1-\rho}{\theta} + \rho + \gamma^2\rho\right\},$$

as $n \to \infty$.

With a power of $1-\beta$, the sample size estimate (n) to test $H_0 : p = p_0$ versus $H_1 : p = p_1$ is

$$n = \frac{(z_{1-\alpha/2} + rz_{1-\beta})^2}{(p_1 - p_0)^2}\left\{\frac{1-\rho}{\theta} + \rho + \gamma^2\rho\right\}p_0(1-p_0), \tag{7}$$

where $r = \sigma_1(T)/\sigma_0(T) = \sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}$. When the number of observations is constant across all subjects, the sample size formula (7) reduces to (6). For a given level of power, the required number of subjects with variable number of observations is always larger than that with equal number of observations across all subjects.

5 An Example

Here we provide, as an example, the sample size estimate for the sensitivity of an enzymatic diagnostic test (2). An enzymatic diagnostic test was employed to decide whether a site was

infected by at least one of two organisms, *treponema denticola* and *bacteroides gingivalis*. Each subject contributed a different number of infected sites which were determined by the gold standard (an antibody assay against the two organisms). Table 1 shows the data of Hujoel *et al.* (2) that contains 29 subjects of which the number of true positive test results (m_i^+) and the number of infected sites (m_i). The distribution of the number of sites can be estimated from the observed distribution of the number of sites. Table 2 gives the observed and the estimated distribution of the number of sites (m).

Suppose we want to use the above data as pilot data to design a similar experiment to test the hypothesis $H_0 : p = 0.7$ versus $H_1 : p = 0.8$. From Table 1, we obtain the intracluster correlation coefficient estimate $\hat{\rho} = 0.2$ from the data using the ANOVA method. From Table 2, we calculate $E(M) = 4.90$, $var(M) = 1.20$, and $\gamma = 0.23$. From Equation (7), the estimated sample sizes required for the experiment are 58 and 75 subjects for 80% and 90% power, respectively. Hujoel *et al.* (2) provided specificity data in the same format as the sensitivity data in Table 1. The sample size estimate for the specificity of an enzymatic diagnostic test can be conducted in the same way as we did for the sensitivity data.

6 Simulation study

We investigated the performance of the sample size formula, Equation (7), for the proportion test through simulation. Since the number of observations per subject is frequently unbalanced in medical studies, we generated the number of observations per subject using the truncated negative binomial distribution below 1 (26), which has probability mass function (27)

$$f(k) = \frac{(s+k-1)!p^s(1-p)^k}{(s-1)!k!(1-p^s)}, \tag{8}$$

which has mean

$$\mu = \frac{s(1-p)}{p(1-p^s)}$$

and variance

$$\sigma^2 = \frac{\mu[1-s(1-p)p^s]}{p(1-p^s)}.$$

The imbalance in the number of observations per subject is measured by the quantity $\kappa = 1/(1 + \sigma^2/\mu^2)$. The smaller κ, the larger the variation in the number of observations per subject. The

number of observations is the same across all subjects when $\kappa = 1$. The number of observations for each subject is generated from the truncated negative binomial distribution with mean of $\mu = 5$, 10, and 20, and the imbalance parameter of $\kappa = 0.6$, 0.8, and 1.0, which corresponds to severe variability, moderate variability, and no variability in the number of observations per subject. We used ρ values of 0.05, 0.1, 0.3, and 0.5. Here, we test the null hypothesis $H_0 : p = p_0$ against the alternative hypothesis $H_1 : p = p_1$ with $\alpha = 0.05$, $1 - \beta = 0.9$ and (p_0, p_1)=(0.6, 0.7), (0.8, 0.9), or (0.7,0.9). The required number of subjects is estimated by Equation (7) for given values of p_0, p_1, ρ, κ, μ, α, and β. The correlated binary data are generated by the method of Lunn and Davies (28) conditional on the number of observations per subject and the estimated number of subjects. The sample size formula, Equation (7), was derived using the weighting method that assigns equal weight to each observation. We compared the performance of the proposed sample size formula with that of the parametric sample size formula assigning equal weight to each observation (10). We conducted 5,000 experiments for each parameter combination. Empirical power was computed as the proportion of 5,000 samples in which the null hypothesis was rejected by the the test statistic (Equation (5)).

Table 3 reports the estimated sample sizes and the corresponding empirical powers for testing $H_0 : p = 0.6$ versus $H_1 : p = 0.7$ at significance level of 0.05 and with a power of 90%. In general, empirical powers are close to the nominal power of 90% for both the proposed method and the parametric method. Table 4 and Table 5 present empirical powers for testing $H_0 : p = 0.8$ versus $H_1 : p = 0.9$ and for testing $H_0 : p = 0.7$ versus $H_1 : p = 0.9$, in which empirical powers using the proposed method are generally closer to the nominal power than the parametric method. It is noticeable that the nonparametric test yields greater power than the parametric test at a cost of more subjects with the exception in one entry in Table 3 and four entries in Table 5. As κ decreases, the number of observations varies more and more severely, the required number of subjects increases. The sample size estimates increase as the intracluster correlation ρ increases. When the mean number of observations μ increases, the required number of subjects decreases. As the specified probability of success p_1 gets closer to p_0, the estimated number of subjects becomes larger.

7 Discussion

We proposed a nonparametric sample size calculation approach for the control of sensitivity and specificity of diagnostic tests with multiple observations per subject. Since the sample size formula is based on the asymptotic theory, simulation studies are conducted to evaluate finite sample performances of the proposed sample size formula. We compared the performance of the proposed sample size formula with that of the parametric sample size formula of Jung *et al.* (10) that assigns equal weight to each observation for the proportion test. Simulation studies show that the proposed sample size formula generally yields empirical powers closer to the nominal level than the parametric method. Simulation studies also show that the number of subjects required increases as the variability in the number of observations increases and the intracluster correlation increases. The test statistic, Equation (5), belongs to the family of cluster-weighted multivariate sign statistics considered by Larocque et. al. (16). In calculation of the test statistic all observations receive equal weights. The effects of different weighting schemes, such as equal weights to subjects and optimal weights, on sample size estimates need to be investigated.

Acknowledgment

This work was supported in part by NIH grants UL1 RR024982, P50CA70907, and NASA NNJ05HD36G.

REFERENCES

1. Fleiss JL, Levin B, Paik CM. *Statistical Methods for rates and proportions.* New York: Wiley, 2003.

2. Hujoel P, Moulton L, Loesche W. Estimation of sensitivity and specify of site-specific diagnostic tests. *Journal of Periodontal Research* 1990;**25**:193-196.

3. Smith P, Hadgu A. Sensitivity and specificity for correlated observations. *Statistics in Medicine* 1992;**11**:1503-1509.

4. Ahn C. An evaluation of methods for the estimation of sensitivity and specificity of site-specific diagnostic tests. *Biometrical Journal* 1997;**7**:793-807.

5. Ahn C. Statistical methods for the estimation of sensitivity and specificity of site-specific diagnostic tests. *Journal of Periodontal Research* 1997;**32**:351-354.

6. Jung SH, Ahn C. Estimation of response probability in correlated binary data: a new approach. *Drug Information Journal* 2000;**34**:599-604.

7. Lee E, Dubin N. Estimation and sample size considerations for clustered binary responses. *Statistics in Medicine* 1994;**13**:1241-1252.

8. Hayes RJ, Bennett R. Sample size calculation for cluster-randomized trials. *International Journal of Epidemiology* 1999;**28**:319-326.

9 Donner A, Klar N. *Design and analysis of cluster randomization trials in health research.* Oxford University Press, 2000.

10. Jung SH, Kang SH, Ahn C. Sample size calculation for clustered binary data. *Statistics in Medicine* 2001;**20**:1971-1982.

11. Braun TM. A mixed model formulation for designing cluster randomized trials with binary outcomes. *Statistical Modelling* 2003;**3**:233-249.

12. Eldridge S, Kerry S, Ashby D. Sample size for cluster randomized trials: effects of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology* 2006;**35**:1292-1300.

13. Liu G, Liang KY. Sample size calculation for studies with correlated observations. *Biometrics* 1997;**53**:937-947.

14. Pan W. Sample size and power calculations with correlated binary data. *Controlled Clinical Trials* 2001;**22**:211-227.

15. Larocque, D. The Wilcoxon signed-rank test for cluster correlated data. *Statistical Modeling and Analysis for Complex Data Problems*, P. Duchesne and B. Rémillard (eds), New York: Springer 2005;309-323.

16. Larocque D, Nevalainen J, Oja H. A weighted multivariate sign test for cluster-correlated data. *Biometrika* 2007;**94**:267-283.

17. Rosner B, Glynn RJ, Lee MLT. The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics* 2006;**62**:185-192.

18. Rosner B, Glynn RJ, Lee MLT. Extension of the rank sum test for clustered data: two-group comparison with group membership defined at the subunit level. *Biometrics* 2006;**62**:1251-1359.

19. Gerard PD, Schucany WR. A enhanced sign test for dependent binary data with small numbers of clusters. *Computational Statistics and Data Analysis* 2007;**51**:4622-4632.

20. Datta S, Satten GA. A signed-rank test for clustered data. *Biometrics* 2008;**64**:501-507.

21. Noether GE. Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* 1987;**82**:645-647.

22. Molenberghs G, Ryan LM. An Exponential family model for clustered multivariate binary data. *Environmetrics* 1999;**10**:279-300.

23. Cox DR, Wermuth N. A note on the quadratic exponential binary distribution. *Biometrika* 1994;**81**:403-408.

24. Donald A, Donner A. A simulation study of the analysis of sets of 2 x 2 contingency tables under cluster sampling: Estimation of a common odds ratio. *J. Amer. Statist. Assoc.* 1990;**85**:537-543.

25. Ridout MS, Demetrio CG, Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999;**55**:137-148.

26. Donner A, Koval J. A procedure for generating group sizes from a one-way classification with a specified degree of imbalance. *Biometrical J* 1987;**29**:181-187.

27. Johnson NL, Kotz S. *Distribution in Statistics. Discrete Distributions.* New York: Wiley, 1969.

28. Lunn AD, Davies SJ. A note on generating correlated binary variables. *Biometrika* 1998;**85**:487-490.

Table 1: Pilot data for sensitivity m_i^+/m_i from $n = 29$ subjects.

3/6, 2/6, 2/4, 5/6, 4/5, 5/5, 4/6, 3/4, 2/4, 3/4, 5/5, 4/4, 6/6, 3/3, 5/6, 1/2, 4/6, 0/4, 5/6, 4/5, 4/6, 0/6, 4/5, 3/5, 0/2, 2/6, 2/4, 5/5, 4/6

Table 2: Distribution of cluster sizes (m_i).

	m_i				
	2	3	4	5	6
Observed distribution	2/29	1/29	7/29	7/29	12/29
Projected distribution	0.05	0.05	0.25	0.25	0.4

Table 3: Empirical powers (%) and sample size estimates in parentheses for testing $H_0 : p = 0.6$ vs. $H_1 : p = 0.7$ with $\alpha = 0.05$ and $\beta = 0.10$ from 5,000 simulations. The sample size refers to the number of clusters.

			Method	
κ	ρ^a	μ^b	ST^c	PT_u^d
0.6	0.05	5	92(66)	91(61)
		10	89(43)	88(40)
		20	89(32)	87(29)
	0.1	5	91(84)	90(77)
		10	89(62)	88(57)
		20	89(51)	87(47)
	0.3	5	90(154)	89(142)
		10	90(137)	89(126)
		20	89(129)	88(119)
	0.5	5	89(224)	89(206)
		10	90(212)	89(195)
		20	89(206)	89(190)
0.8	0.05	5	89(61)	88(56)
		10	89(38)	88(35)
		20	89(27)	87(25)
	0.1	5	90(74)	88(68)
		10	89(52)	89(48)
		20	89(41)	87(38)
	0.3	5	90(124)	89(114)
		10	90(107)	89(99)
		20	89(99)	88(91)
	0.5	5	90(174)	88(160)
		10	88(162)	89(149)
		20	90(156)	89(144)
1	0.05	5	89(58)	88(53)
		10	90(35)	87(32)
		20	89(24)	88(22)
	0.1	5	90(68)	88(62)
		10	89(46)	87(42)
		20	89(35)	87(32)
	0.3	5	90(106)	89(98)
		10	90(89)	89(82)
		20	90(81)	88(74)
	0.5	5	90(144)	88(133)
		10	90(132)	89(122)
		20	90(126)	88(116)

a: ρ is an intracluster correlation coefficient

b: μ is the mean cluster size of a truncated negative binomial distribution below 1

c: Sign test for clustered binary data

d: Parametric test assigning equal weight to each site from Jung *et al.* (2001)

Table 4: Empirical powers (%) and sample size estimates in parentheses for testing $H_0 : p = 0.8$ vs. $H_1 : p = 0.9$ with $\alpha = 0.05$ and $\beta = 0.10$ from 5,000 simulations. The sample size refers to the number of clusters.

			Method	
κ	ρ^a	μ^b	ST^c	PT_u^d
0.6	0.05	5	90(38)	85(26)
		10	85(25)	83(17)
		20	84(18)	83(13)
	0.1	5	88(48)	85(33)
		10	86(36)	83(25)
		20	84(29)	82(21)
	0.3	5	88(88)	84(61)
		10	88(78)	83(54)
		20	88(74)	83(51)
	0.5	5	89(128)	84(89)
		10	88(121)	84(84)
		20	88(118)	83(82)
0.8	0.05	5	89(35)	83(24)
		10	87(22)	83(15)
		20	85(16)	83(11)
	0.1	5	88(42)	83(29)
		10	85(30)	82(21)
		20	84(24)	81(17)
	0.3	5	88(71)	84(49)
		10	87(61)	83(43)
		20	85(56)	82(39)
	0.5	5	88(99)	82(69)
		10	88(93)	83(64)
		20	88(89)	85(62)
1	0.05	5	88(33)	83(23)
		10	87(20)	83(14)
		20	86(14)	83(10)
	0.1	5	87(39)	84(27)
		10	85(26)	82(18)
		20	83(20)	81(14)
	0.3	5	88(61)	84(42)
		10	87(51)	82(35)
		20	86(46)	84(32)
	0.5	5	88(82)	84(57)
		10	89(76)	84(53)
		20	88(72)	83(50)

a: ρ is an intracluster correlation coefficient

b: μ is the mean cluster size of a truncated negative binomial distribution below 1

c: Sign test for clustered binary data

d: Parametric test assigning equal weight to each site from Jung *et al.* (2001)

Table 5: Empirical powers (%) and sample size estimates in parentheses for testing $H_0 : p = 0.7$ vs. $H_1 : p = 0.9$ with $\alpha = 0.05$ and $\beta = 0.10$ from 5,000 simulations. The sample size refers to the number of clusters.

			Method	
κ	ρ^a	μ^b	ST^c	PT_u^d
0.6	0.05	5	84(12)	78(7)
		10	84(8)	81(5)
		20	83(6)	84(4)
	0.1	5	86(15)	82(9)
		10	85(11)	83(7)
		20	82(9)	84(6)
	0.3	5	87(27)	81(16)
		10	85(24)	81(14)
		20	86(23)	80(13)
	0.5	5	88(39)	82(23)
		10	87(37)	81(21)
		20	87(36)	81(21)
0.8	0.05	5	85(11)	73(6)
		10	85(7)	79(4)
		20	84(5)	82(3)
	0.1	5	85(13)	79(8)
		10	83(9)	82(6)
		20	80(7)	82(5)
	0.3	5	87(22)	80(13)
		10	85(19)	80(11)
		20	84(17)	80(10)
	0.5	5	87(30)	81(18)
		10	88(28)	79(16)
		20	86(27)	80(16)
1	0.05	5	83(10)	73(6)
		10	85(6)	79(4)
		20	87(5)	84(3)
	0.1	5	85(12)	76(7)
		10	84(8)	80(5)
		20	78(6)	80(4)
	0.3	5	88(19)	79(11)
		10	85(16)	78(9)
		20	83(14)	79(8)
	0.5	5	87(25)	78(15)
		10	87(23)	81(14)
		20	86(22)	82(13)

a: ρ is an intracluster correlation coefficient

b: μ is the mean cluster size of a truncated negative binomial distribution below 1

c: Sign test for clustered binary data

d: Parametric test assigning equal weight to each site from Jung *et al.* (2001)