Journal of Clinical Epidemiology 60 (2007) 250–255

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

A cluster-adjusted sample size algorithm for proportions was developed using a beta-binomial model

G.T. Fosgate*

Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX 77843-4458, USA

Accepted 2 June 2006

Abstract

Objective: The objective of the paper was to design a computer algorithm to calculate sample sizes for estimating proportions incorporating clustered sampling units using a beta-binomial model when information concerning the intraclass correlation is not available.

Study Design and Setting: A computer algorithm was written in FORTRAN and evaluated for a hypothetical sample size situation.

Results: The developed algorithm was able to incorporate clustering in estimated sample sizes through the specification of a beta distribution to account for within-cluster correlation. In a hypothetical example, the usual normal approximation method for estimation of a proportion ignoring the clustered sampling design resulted in a calculated sample size of 107, whereas the developed algorithm suggested that 208 sampling units would be necessary.

Conclusion: It is important to incorporate cluster adjustment in sample size calculations when designing epidemiologic studies for estimation of disease burden and other population proportions in the situation of correlated data even when information concerning the intraclass correlation is not available. Beta-binomial models can be used to account for clustering, and design effects can be estimated by generating beta distributions that encompass within-cluster correlation.

Keywords: Sample size; Proportion; Binomial; Beta-binomial; Cluster; Intraclass correlation

1. Introduction

Sample size calculation is an important design aspect for studies that aim to estimate population proportions including disease prevalence. It is a common situation in human and veterinary medicine that sampling units share many risk factors for disease when they are clustered in time and space. The impact of such clustering should be incorporated in variance estimates when calculating confidence intervals. It would be inappropriate to calculate sample size estimates under the independence assumption if it is known in advance that sampling units are not independent due to clustering.

Sample size estimates are often adjusted for clustering by incorporating an inflation factor that is often referred to as the design effect (DE) [1–3] or the variance inflation factor [2,4,5]. Results of the usual sample size formulas are simply increased by this factor to account for correlation in sampling units that will result due to the sampling technique. Often, data necessary for estimation of the DE are not available before performing the study. It therefore becomes difficult, if not impossible, to adjust sample size estimates without first performing a preliminary investigation to estimate the intraclass correlation. This correlation is defined as the proportion of total variation within sampling units that can be accounted for by variation among clusters [1–3,6].

Correlated sampling units in the situation of disease burden studies imply that the prevalence of a particular outcome is not equal among clusters due to factors in common that increase risk (or duration of the outcome). Within each cluster, the probability of the outcome is assumed to be equal for all individuals. However, this probability is expected to be different among clusters in an amount dependent upon the intraclass correlation. A convenient model for these data (correlated binary outcomes) is the beta-binomial model [4,7,8]. This model has a hierarchical structure with the first level modeling disease within each cluster as binomial counts. The second level models the probability of disease among clusters as a beta distribution. The objective of the paper reported here was to design a simple computer algorithm to calculate sample sizes for estimating proportions incorporating clustered sampling units using a beta-binomial model when information concerning the intraclass correlation is not available.

* Corresponding author. Tel.: 979-845-3203; fax: 979-847-8981.
E-mail address: gfosgate@cvm.tamu.edu (G.T. Fosgate).

0895-4356/07/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2006.06.010

2. Methods

2.1. Sample size example

A clinical application of this cluster-adjusted sample size routine would be estimation of an overall proportion of health care–associated infections (HAIs) across hospital wards. A hospital administrator might be interested in estimating the overall probability of HAI in the entire institution for reporting purposes. Baseline risk of HAI development is expected to vary depending upon hospital ward because of the distribution of illnesses and patient demographics. A random sampling approach stratified on hospital ward would be the most efficient design to use because within-strata variation is expected to be less than among-strata variation. A simple binomial model for calculation of the necessary sample size would be inappropriate because it cannot account for stratification and different baseline risks for HAI. The beta-binomial models the variation in baseline HAI risk across strata using a beta distribution allowing for adjustment of the sample size for inherent clustering. The sample size routine to solve this problem requires the following inputs: expected mean HAI proportion over all clusters, desired error limit of this proportion, desired level of confidence, number of clusters to be sampled, and either the 5th or the 95th percentile for the beta distribution modeling the proportion among clusters.

2.2. Beta-binomial distribution

The sample size model defines k as the number of distinct groups (strata) of study individuals. The ith group contains m_i individuals, each having a dichotomous response X_{ij} ($i = 1 \ldots k$; $j = 1 \ldots m_i$). The two values of X_{ij} are 0 (failure; outcome not present) and 1 (success; outcome present). We calculate $Y_i = \Sigma_j X_{ij}$ as the total number of successes in the ith group. The sample size routine assumes that individuals are independent and that the probability of success varies from group to group but not among individuals within the same group.

The beta-binomial model is derived as a mixture distribution where the probability of success varies from group to group according to a beta distribution with parameters α and β, and conditional on this probability, Y_i has a binomial distribution. The data model is therefore specified as

$$\begin{aligned} &Y_i \sim \text{Binomial}\,(m_i, \pi_i),\ \ i = 1, 2, \ldots k \\ &\pi_i \sim \text{Beta}\,(\alpha, \beta) \\ &Y_i \sim \text{Beta-binomial}\,(m_i, \alpha, \beta) \end{aligned}$$

where m_i is the cluster sample size and k is the number of clusters. The specific model used further assumes that all m_i are equal and therefore can be denoted as a common group size, m. The mean, μ, or marginal probability of success is equal to $\alpha/(\alpha + \beta)$.

2.3. Cluster-level variability specification

The beta distribution used to account for expected variation of the proportion across clusters is found by eliciting investigator opinion. The investigator is asked to provide the expected overall proportion, which is considered the mean of the beta distribution. If this mean is $\leqslant 0.5$, then the investigator is asked to provide the expected 95th percentile of the distribution. The clusters are believed to have a proportion less than this value with 95% probability based on the investigator's biologic understanding of the problem. If the mean is > 0.5, then the investigator is asked to provide the expected 5th percentile of the distribution; the clusters are believed to have a proportion less than this with only 5% probability. The requested percentile is to elicit the investigator's opinion concerning expected variation of the proportion across study clusters. The beta distribution is generated through the specification of the expected overall proportion and values associated with the 5th or 95th percentile. For purposes of this discussion, precision of the beta distribution is defined as the length measured from the mean (expected overall proportion) to the specified percentile limit. The assumption is then made that the beta distribution over this interval can be well approximated by a segment of a normal curve. This simplifying assumption allows for determination of beta parameters (α, β) using the standard formula for calculating a confidence interval [9]:

$$\text{Confidence interval} = \text{point estimate} \pm (\text{standard error} \times \text{reliability coefficient}).$$

The precision, as previously defined, is the same as half the length of a standard confidence interval or simply the standard error multiplied by the reliability coefficient (RC). The reliability coefficient, in this case, is the standard normal z score for a 90% confidence interval, which is 1.6449. The standard deviation of a beta distribution replaces the standard error in the formula for the confidence interval. The formula for the standard deviation [10–12] is

$$\sigma = \sqrt{\frac{\alpha \times \beta}{(\alpha + \beta)^2 \times (\alpha + \beta + 1)}}$$

which can be algebraically simplified as

$$\sigma = \sqrt{\frac{\mu(1 - \mu)}{\alpha + \beta + 1}}$$

where μ is the mean of the beta distribution. The formula for the standard deviation can be substituted into the equation for the precision and rewritten to solve for the sum of the beta parameters as

$$\alpha + \beta = \frac{\mu(1 - \mu)}{(\text{precision}/RC)^2} - 1.$$

The specified proportion is considered the mean of the beta distribution, and this allows for the calculation of each parameter, and therefore complete specification of the distribution, by using the following relationship: $\mu = \alpha/(\alpha + \beta)$. Commercially available software [13] was used to compare distribution inputs with actual cumulative probability intervals of the generated beta distributions.

2.4. DE estimation

DE can be defined as the variance of the sampling design compared to simple random sampling [4]. It is used to adjust sample size calculations that were made based on the assumption of simple random sampling [1] and is calculated [2,6] as

$$DE = 1 + \rho(m - 1) \tag{1}$$

where ρ is the intraclass correlation and m is the sample size within each cluster.

The beta-binomial assumption allows calculation of the intraclass correlation from the generated beta distribution [7,8] as

$$\rho = \frac{1/(\alpha + \beta)}{1 + 1/(\alpha + \beta)}.$$

The effective sample size (ESS) is the number of uncorrelated (not clustered) sampling units that would yield an amount of statistical information equivalent to that provided by the clustered data. ESS can be calculated [3] as

$$ESS = mk/DE$$

where m is the cluster sample size and k the number of clusters. These relationships can be used to algebraically solve for m, which is written as

$$m = \frac{ESS - \rho(ESS)}{k - \rho(ESS)}. \tag{2}$$

The cluster sample size (samples collected per cluster) and therefore the total sample size are found by providing the number of clusters (k) and ESS, and using the intraclass correlation determined from the generated beta distribution. The assumption of a fixed number of clusters of equal size provides the necessary information for calculation of the DE as shown in equation (1). The inherent limitation of the denominator of equation (2) to be positive was evaluated graphically for a hypothetical sample size situation.

2.5. Algorithm development

The cluster sample size algorithm was written in FORTRAN [14] incorporating a previously published modified exact sample size routine [15] using a commercially available software development platform [16]. The inputs of the algorithm are the hypothesized proportion (mean over all clusters), desired error limit, desired level of confidence, number of clusters to be sampled, and either the 5th or the 95th percentile for the beta distribution modeling the proportion among clusters. The modified exact sample size based on the previously published algorithm is calculated along with the usual normal approximation method using a standard formula [17] for comparison. Parameters of the beta distribution are generated leading to estimation of the DE, and the modified exact sample size is used as the ESS, or the number of sampling units that would be necessary under the assumption of independence. The cluster-adjusted sample size is then calculated as the DE multiplied by the modified exact sample size. The FORTRAN program code was compiled into a DOS-based application and is available by contacting the author or from the online version of the article from the journal's website at www.elsevier.com.

3. Results

3.1. Beta distribution and DE estimation

Beta distributions estimated by the algorithm had 95% cumulative probability distributions similar to specified values over the range of evaluated proportions (Table 1). Deviations from the true limits were recognized for proportions ≤0.2 and became larger as the proportion decreased toward zero. DE estimations varied depending upon the 95th percentile inputs, hypothesized proportion, and evaluated number of clusters (Table 2). Larger DEs were observed with beta distributions representing larger variation (larger distance from proportion to 95th percentile). Larger DEs were also observed with smaller number of clusters for sampling. Estimated DEs tended to increase closer to the boundary value of zero. Increasing the sample size per cluster had a limited effect on increasing the ESS for a fixed

Table 1
True characteristics of beta distributions generated by the algorithm to approximate variation in proportions among clusters

Proportion	95th percentile input	Estimated beta parameters	Distribution characteristics[a]	
			Mean	95th percentile
0.50	0.52	845.03, 845.03	0.50	0.52
	0.55	134.78, 134.78	0.50	0.55
	0.60	33.32, 33.32	0.50	0.60
0.40	0.42	648.97, 973.45	0.40	0.42
	0.45	103.50, 155.25	0.40	0.45
	0.50	25.57, 38.36	0.40	0.50
0.30	0.32	425.85, 993.65	0.30	0.32
	0.35	67.88, 158.40	0.30	0.35
	0.40	16.75, 39.07	0.30	0.40
0.20	0.22	216.26, 865.02	0.20	0.22
	0.25	34.43, 137.73	0.20	0.25
	0.30	8.46, 33.83	0.20	0.31
0.10	0.12	60.78, 547.00	0.10	0.12
	0.15	9.64, 86.76	0.10	0.15
	0.20	2.34, 21.02	0.10	0.22
0.05	0.07	16.02, 304.29	0.05	0.07
	0.10	2.52, 47.89	0.05	0.11
	0.15	0.59, 11.26	0.05	0.18

[a] True mean and 95th percentile of induced beta distribution.

Table 2
DE estimation for a number of proportions, clusters, and beta distributions for sample size estimation assuming an error limit of 0.05 and 95% confidence

		Beta 1[b]				Beta 2[b]				Beta 3[b]			
	Uncorrected	No. of clusters				No. of clusters				No. of clusters			
Proportion	sample size[a]	2	4	8	16	2	4	8	16	2	4	8	16
0.50	384	1.1	1.1	1.0	1.0	3.4	1.5	1.2	1.1	NS	NS	3.4	1.5
0.45	380	1.1	1.1	1.0	1.0	3.4	1.5	1.2	1.1	NS	NS	3.4	1.5
0.40	369	1.1	1.1	1.0	1.0	3.4	1.5	1.2	1.1	NS	NS	3.4	1.5
0.35	351	1.1	1.1	1.0	1.0	3.4	1.5	1.2	1.1	NS	NS	3.4	1.5
0.30	323	1.1	1.1	1.0	1.0	3.4	1.5	1.2	1.1	NS	NS	3.4	1.5
0.25	288	1.1	1.1	1.0	1.0	3.4	1.5	1.2	1.1	NS	NS	3.4	1.5
0.20	249	1.1	1.1	1.0	1.0	3.5	1.6	1.2	1.1	NS	NS	3.5	1.5
0.15	200	1.1	1.1	1.0	1.0	3.6	1.6	1.2	1.1	NS	NS	3.5	1.5
0.10	148	1.1	1.1	1.0	1.0	4.1	1.6	1.2	1.1	NS	NS	4.0	1.5
0.05	97	1.2	1.1	1.0	1.0	17.3	1.9	1.3	1.1	NS	NS	16.3	1.7

NS = not possible to solve.

[a] Sample size based on mid-*P* modified exact methods.

[b] Induced beta distribution calculated for specified proportion with 95th percentile = proportion + 0.02 (Beta 1), proportion + 0.05 (Beta 2), proportion + 0.10 (Beta 3) similar to data presented in Table 1.

value of intraclass correlation and number of clusters (Fig. 1).

3.2. Clinical example using HAI

HAIs are localized or systemic infections (or diseases secondary to toxins produced by infectious agents) not present (or incubating) at the time of admission but acquired by a patient in health care settings [18]. These infections have current importance because they contribute to a death every 6 minutes and cost the health care system U.S. $4.5 billion based on 1995 data [19]. Four U.S. states have already enacted legislation requiring health care organizations to disclose HAI rates to allow consumers to make more informed health care decisions [18].

For this example, we will assume that a hypothetical hospital comprises eight wards: critical care, pediatric, maternity, geriatric, surgical, cardiac, respiratory, and general medicine. The hospital administration is interested in estimating the overall incidence of HAI (proportion per hospital stay) including follow-up of patients after discharge. It is expected that the risk of acquiring an HAI is different for each ward, and therefore a stratified random sample of patients (beds) from all wards will be used. The degree of dependency in the proportion of HAI due to hospital ward is currently unknown. However, in developed countries it has been estimated that as many as 5%−10% of all hospitalized patients acquire an HAI [20].

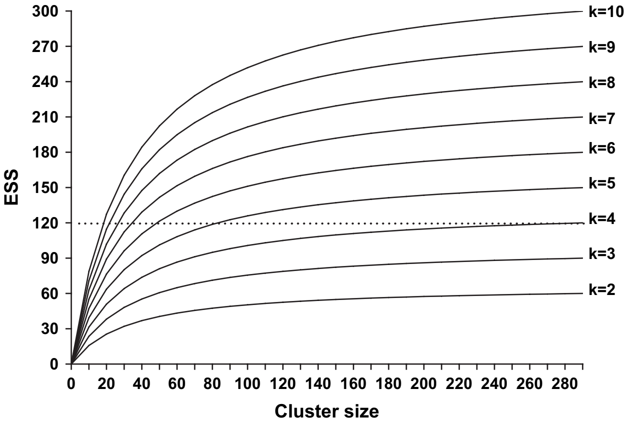


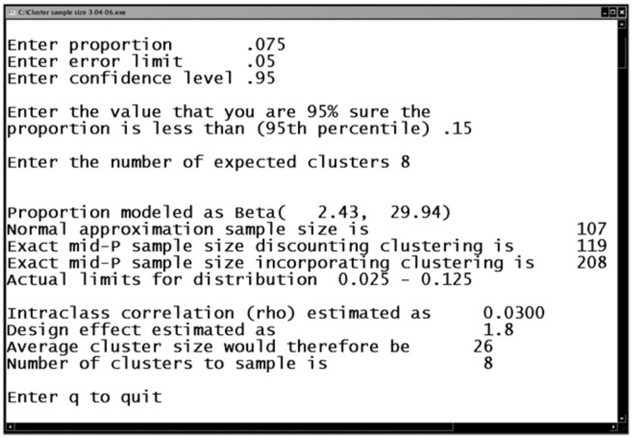
Fig. 1. The consequence of increasing cluster size on ESS for several numbers of clusters (*k*). Calculations are based on an intraclass correlation of 0.03, and the ESS of 119, which corresponds to estimating a proportion of 0.075 with an error of 0.05 and 95% confidence, is denoted by the dashed line.

To calculate the necessary sample size to estimate the proportion of HAI in this hypothetical hospital, we will assume a most likely proportion of 0.075 (middle of reported range) and a 95th percentile of 15% (double the overall expected). We calculate the sample size necessary to estimate this proportion with an acceptable error of ±0.05 at the 95% level of confidence. The total sample size for this situation is therefore estimated as 208 patients overall, with 26 sampled from each ward (Fig. 2). The usual normal approximation method discounting clustering estimated a total sample size of 107 patients.

4. Discussion

Epidemiologic studies for investigation of disease burden require sample size calculations that will yield estimated proportions with the desired level of precision. The method of calculation should therefore be appropriate for the expected statistical analysis. People and animals are often clustered in management units with many characteristics in common resulting in dependence, or correlation, related to the outcome of interest. This correlation among sampling units should be accounted for during both the design and analysis phases of the study.

The beta-binomial distribution is a convenient model for correlated binary outcomes. Specification of the beta



```
Enter proportion       .075
Enter error limit      .05
Enter confidence level .95

Enter the value that you are 95% sure the
proportion is less than (95th percentile) .15

Enter the number of expected clusters 8


Proportion modeled as Beta(   2.43,  29.94)
Normal approximation sample size is                      107
Exact mid-P sample size discounting clustering is        119
Exact mid-P sample size incorporating clustering is      208
Actual limits for distribution  0.025 - 0.125

Intraclass correlation (rho) estimated as     0.0300
Design effect estimated as                    1.8
Average cluster size would therefore be      26
Number of clusters to sample is               8

Enter q to quit
```

Fig. 2. Screen capture of the cluster-adjusted sample size algorithm to calculate the number of patients necessary to sample in the hypothetical HAI study.

distribution based on the investigator's knowledge of the expected percentile limits allows for estimation of the intraclass correlation without performing a pilot study. In the present study, beta distributions generated in this manner tended to have actual cumulative probability distributions similar to the specified levels except for evaluated proportions close to the boundary value of zero. Beta distributions are frequently skewed, and the magnitude depends upon the location of the mode and the standard deviation. Highly skewed beta distributions will not be well approximated by a normal curve, and this is the reason for discrepancies in the desired and actual limits in certain situations.

Parameters (α, β) of the generated beta distributions provide information for estimating the intraclass correlation in the beta-binomial model. DEs are then estimated from the intraclass coefficient. The DE is calculated assuming a fixed number of clusters of equal size. The observation that the DE increased as the proportion approached zero is likely due to the fact that generated beta distributions had smaller α and β parameters resulting in larger values for the intraclass correlation. DEs, however, cannot be estimated for all sample size situations because it is necessary for $k > \rho \times \text{ESS}$ in the formula for calculating the cluster size (m). With a fixed number of clusters (k), the maximum attainable ESS would be limited by the intraclass correlation (ρ) and can be mathematically shown to be equal to k/ρ. Another approach that could be followed would be to fix the cluster size as an input in the algorithm and then algebraically solve for the number of clusters to sample. The author felt that the number of clusters to sample was a more practical input for epidemiologists designing studies for the evaluation of disease burden.

A drawback to this approach is the assumption of equal cluster sizes because it is not possible to relax this requirement in the current form of the algorithm. Other limitations that could be addressed in future modifications are the assumption that the outcome is determined in sampled individuals without error (sensitivity = specificity = 1) and that the sample size per cluster is relatively small compared to the total cluster size (binomial assumption). The algorithm also lacks a menu-driven interface because it was compiled as a simple DOS-based program.

The cluster sample size algorithm was written in FORTRAN because the mid-P exact sample size routine for the binomial component of the model was previously written in this computing language. The binomial probability function is computationally intense, and this prevents the sample size algorithm from being able to solve all possible combinations of proportion, interval width, and confidence level. The program will fail when individual binomial probabilities become functionally zero (calculated as zero due to precision limitations) and will not report a sample size in those instances. The computational intensity might also cause the algorithm to run slowly on computers with slower processor speeds.

The algorithm was described for the design of a disease burden study, but will function equally well for estimation of sample sizes associated with any population proportion and not simply those <0.5. The symmetry in the binomial distribution means that a sample size calculated for proportion, p, will be the same as would be estimated for $1 - p$ and therefore only proportions <0.5 were presented. The purpose of the designed algorithm was to allow for incorporation of cluster adjustment in estimation of sample sizes for proportions when no information is available concerning the intraclass correlation. The procedure for generating the beta distribution to allow for this adjustment is similar to what is employed to elicit prior probabilities concerning expert opinion in Bayesian analyses [12,21]. It would be theoretically possible to elicit distributions based on exact cumulative beta probability functions, but this would require an iterative procedure and more computing resources. The normal approximation method used was accurate for most evaluated situations, and it is unclear if exact methods would lead to better approximations. Exact methods could actually be less precise depending upon the tolerance level set by the programmer. A software routine, Betabuster, has been developed for generating beta distributions based on the mode and percentile limits and is available at http://www.epi.ucdavis.edu/diagnostictests/ under the software category. The specific method used for producing the beta distributions is not described, but the resulting distributions have deviations from specified inputs similar to those obtained from the method described here.

5. Conclusion

It is important to incorporate cluster adjustment in sample size calculations when designing epidemiologic studies for estimation of disease burden and other population proportions in the situation of correlated data. Beta-binomial models can be used to account for clustering, and DEs can be estimated by generating beta distributions that encompass within-cluster correlation.

Acknowledgments

I would like to thank the anonymous referees for their helpful suggestions, which resulted in a better overall manuscript.

References

[1] Campbell MK, Mollison J, Grimshaw JM. Cluster trials in implementation research: estimation of intracluster correlation coefficients and sample size. Stat Med 2001;20:391–9.

[2] Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. Stat Med 2002;21:3757–74.

[3] Killip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. Ann Fam Med 2005;2:204–8.

[4] Bohning D, Greiner M. Prevalence estimation under heterogeneity in the example of bovine trypanosomosis in Uganda. Prev Vet Med 1998;36:11–23.

[5] Fleiss JL, Levin BA, Paik MC. Statistical methods for rates and proportions. 3rd ed. Hoboken, NY: John Wiley; 2003. p. 213.

[6] McDermott JJ, Schukken YH, Shoukri MM. Study design and analytic methods for data collected from clusters of animals. Prev Vet Med 1994;18:175–91.

[7] Donner A, Donald A. The statistical analysis of multiple binary measurements. J Clin Epidemiol 1988;41:899–905.

[8] Ridout MS, Demetrio CG, Firth D. Estimating intraclass correlation for binary data. Biometrics 1999;55:137–48.

[9] Daniel WW. Biostatistics: a foundation for analysis in the health sciences. 7th ed. New York: Wiley; 1999. p. 156.

[10] Aitken CG. Sampling—how big a sample? J Forensic Sci 1999;44: 750–60.

[11] Agresti A. Categorical data analysis. 2nd ed. New York: Wiley-Interscience; 2002. p. 606.

[12] Suess EA, Gardner IA, Johnson WO. Hierarchical Bayesian model for prevalence inferences and determination of a country's status for an animal pathogen. Prev Vet Med 2002;55:155–71.

[13] @Risk, version 4.5.2. Ithaca, NY: Palisade Corporation; 2002.

[14] International Business Machines Corporation. Programming Research Group. Preliminary report: specifications for the IBM Mathematical FORmula TRANslating System, FORTRAN. New York: The Corporation; 1954.

[15] Fosgate GT. Modified exact sample size for a binomial proportion with special emphasis on diagnostic test parameter estimation. Stat Med 2005;24:2857–66.

[16] Compaq visual Fortran: professional edition, version 6.6. Palo Alto, CA: Hewlett-Packard Company; 2000.

[17] Thrusfield M. Veterinary epidemiology. 3rd ed. Ames, IA: Blackwell Publishing; 2005. pp. 232–3.

[18] McKibben L, Horan T, Tokars JI, Fowler G, Cardo DM, Pearson ML, et al. Guidance on public reporting of healthcare-associated infections: recommendations of the healthcare infection control advisory committee. Am J Infect Control 2005;33:217–26.

[19] Weinstein RA. Nosocomial infection update. Emerg Infect Dis 1998;4:416–20.

[20] Exner M, Kramer A, Lajoie L, Gebel J, Engelhart S, Hartemann P. Prevention and control of health care-associated waterborne infections in health care facilities. Am J Infect Control 2005;33(Suppl I): S26–40.

[21] Branscum AJ, Gardner IA, Johnson WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Prev Vet Med 2005;68:145–63.