Research/Recherche

Efficiency of EPI cluster sampling for assessing diarrhoea and dysentery prevalence

S.S. Yoon,[1] J. Katz,[2] K. Brendel,[3] & K.P. West Jr[2]

This study examines the efficiency of EPI cluster sampling in assessing the prevalence of diarrhoea and dysentery. A computer was used to simulate fieldwork carried out by a survey taker. The bias and variance of prevalence estimates obtained using EPI cluster sampling were compared with those obtained using simple random sampling and cluster (stratified random) sampling. Efficiency ratios, calculated as the mean square error divided by total distance travelled, were used to compare EPI cluster sampling to simple random sampling and standard cluster sampling. EPI cluster sampling may be an appropriate low-cost tool for monitoring trends in the prevalence of diarrhoea and dysentery over time. However, it should be used with caution when estimating the prevalence of diarrhoea at a single point in time because of the bias associated with this cluster sampling method

Introduction

Diarrhoea is one of the major causes of morbidity and mortality among children in developing countries. It has been estimated that over 12000 persons die each day from diarrhoeal disease in developing countries (*1*, *2*). Diarrhoea has been shown to affect the growth of children and is an important cause of malnutrition which can adversely affect child mortality (*3*, *4*). The infectious nature of diarrhoea suggests that faecal contamination of water and food contributes to its transmission (*5*). For example, a contaminated source of water, if serving several families, can cause a cluster of illness.

The prevalence of diarrhoea and dysentery in a community can be estimated through the use of a sample survey. Depending on the method used, a sample survey can give an accurate overview of the disease process within a community. For infectious diseases such as diarrhoea and dysentery, a properly designed and conducted sample survey can provide an epidemiological snapshot of the disease. A diarrhoeal disease survey may also be used to assess the nutritional status of a community, since children affected with diarrhoea and dysentery are often nutritionally deficient. A survey can also establish baseline information prior to an intervention. Repeated over time, cross-sectional surveys are useful surveillance tools for monitoring changes in a population.

There are a number of sampling strategies that can be used to ascertain the prevalence of diarrhoea and dysentery in a community. Theoretically the simplest, but often the most difficult to carry out, is simple random sampling (SRS). SRS requires each member of the population to have a known and equal probability of selection. The list from which the sample is drawn is assumed to include every eligible member. The compilation of this list, often called the sampling frame, is difficult and sometimes impossible, even in situations with large amounts of resources. In developing countries, acquisition of the sampling frame can be especially difficult because of inaccuracies in the existing census, to which are added high migration and birth rates and severely limited resources.

An additional difficulty with SRS which makes it impractical in most situations is that subjects selected in the sample can be geographically spread out. Visiting these subjects can divert resources from other areas of the survey or even other health pro-

[1] Epidemic Intelligence Service Officer, National Center for Environmental Health, Centers for Disease Control and Prevention (CDC), Mail Stop F-47, 4770 Buford Highway NE, Atlanta, GA 30341-3724, USA. Dr Yoon was not affiliated with CDC when this work was carried out. Requests for reprints should be sent to Dr Yoon.

[2] Associate Professor, School of Hygiene and Public Health, Johns Hopkins University, Baltimore, MD, USA.

[3] Computer consultant, Atlanta, GA, USA.

Reprint No. **5796**

grammes. Inaccessible areas containing only one or two subjects may need to be visited. Consequently, SRS is infrequently used.

A method that eliminates many of the logistical problems of SRS is cluster sampling. The primary sampling units are clusters, often defined so as to be convenient for the researcher (*6*). For example, clusters may be city blocks if the survey is conducted in an urban setting, or villages if the survey is conducted in rural areas. For cluster sampling, a complete list of clusters is necessary, which is usually much easier to obtain than the sort of list required for SRS. However, a complete list of individual elements is necessary for selected clusters. SRS is often used for sampling within clusters. If clusters are "large", the amount of resources necessary to list their elements can still be substantial.

Thus, a simplified low-cost methodology (EPI cluster sampling) was developed by the WHO Expanded Programme on Immunization (EPI) and has been used extensively (*7*). Typically, EPI cluster surveys have consisted of 30 clusters of 7 children each (for immunization coverage surveys) (*8*). However, the method has been modified and adapted for other purposes: 30 clusters of 14 children have been used for surveys of immunization coverage, oral rehydration salt availability, and breast-feeding and dietary information (*9*); 30 clusters of 68 children were used in Guinea to monitor selective components of primary health care (*10*); 30 clusters of 70 children were used for surveying neonatal tetanus mortality in Kenya (*11*); 30 clusters of 215, for assessing morbidity and mortality due to diarrhoea in the Central African Republic (*12*); 19 clusters of 10–13 children each, in the Philippines (*13*); and 45 clusters of 8 children, for breast-feeding practices (*14*).

Standard cluster sampling uses SRS for sampling within clusters; however, EPI cluster sampling uses a convenient sampling procedure instead of probability-based SRS for second-stage sampling. Thus, although EPI cluster sampling is easier and less costly, the bias and variance of the estimates obtained can increase.

The present study investigates the cost of conducting EPI cluster sample surveys (expressed as distance travelled to visit the sample subjects) and the bias and variance of the estimates derived with this sampling design.

Materials and methods

Data were derived from a survey of preschool-child morbidity that was carried out prior to a randomized vitamin A supplementation trial in the Terai district of Sarlahi, Nepal (*15*, *16*). Communities enrolled in the trial were mapped, and each house was numbered. We used data from a subset of 40 wards in 25 villages where morbidity histories, anthropometric measurements, and ocular health indicators were collected at the beginning of and at regular intervals during the trial. The 40 wards were selected with probability proportional to size from among 261 wards in the trial.

All children under 5 years of age were identified by a house-to-house census and enrolled in the study. At the outset, respondents were queried to determine if enrolled children had either a 7-day history of diarrhoea (defined as four or more loose, watery stools per day) or dysentery (defined as the presence of blood in the stool regardless of whether diarrhoea was present). Other information on health and nutritional status was also collected. Interviews were conducted by persons with at least 10 years of formal education, but without medical training. Local terms for symptoms were used, which had been previously identified through focus-group discussions within the communities. Children with symptoms on one or more days were considered to be ill. Children not under direct observation by a parent for the past week were considered to have missing morbidity data.

Maps of each of the wards were established for the census. Each house was identified on the maps with the same identifier as used in the database. Although not drawn to scale, the maps were the means of locating households for subsequent visits. The quality of the data collected has been described (*17*). The study population consisted of 4297 children who participated in the baseline survey and had a history of morbidity.

Three different sampling schemes were simulated: SRS, stratified random sampling, and EPI cluster sampling. Modified versions of the EPI cluster sampling scheme were also simulated. The SRS method randomly selected a sample of fixed size from the total population of eligible children, regardless of ward of residence. For stratified random sampling, a fixed number of children in each of the 40 wards was randomly selected. For EPI cluster sampling a fixed number of children in each ward was selected using an EPI-recommended sampling strategy. All simulations were carried out using the established maps.

Like cluster sampling, EPI cluster sampling often uses probability-proportional-to-size (PPS) sampling to select clusters. However, at the second stage, EPI cluster sampling selects a starting household in each community by locating the ward's centre, randomly selecting a direction, and randomly selecting a house from a list of all houses falling along the line drawn from the ward centre to the

periphery in the chosen direction. All eligible children in the selected households are included. If additional children are required, the closest house to the right is visited, until the required number of children are sampled. If a household has more children than required for a given sample size, all children in the household are nevertheless sampled. A version of EPI cluster sampling that has the effect of spreading out the sampled children across the community by visiting the *n*th nearest house to the right was also simulated.

A total of 1000 simulations were performed for each sample size investigated. For SRS, sample sizes of 280, 400, 600, 800, and 1000 were simulated. For stratified random sampling, sample sizes of 7, 10, 15, 20, and 25 children per ward were simulated for all 40 wards, resulting in sample sizes of 280, 400, 600, 800, and 1000. For EPI cluster sampling, a minimum of 7, 10, 15, 20, and 25 children per ward were sampled in each of the 40 wards, resulting in sample sizes of approximately 280, 400, 600, 800, and 1000. Sample sizes were selected to provide a range of estimates of bias and variance for each sampling method and to allow comparisons with previous computer simulations of EPI cluster sampling (*18*, *19*).

Simulations required computer maps representing the locations of households and allowing calculations to be made of the distance travelled in wards for each sampling method. A digitizing tablet and MapInfo (MapInfo Corporation, NY, USA) geographic-information-system software were used to enter household locations, which were referenced to an arbitrary point of origin. The centre of each ward was calculated by taking the average of the coordinates of the households within a ward. For SRS and stratified random sampling, a Turbo Pascal (Borland Corporation, CA, USA) program determined, based on the list of selected children, the child living closest to the ward centre. From this initial household, the nearest household containing a selected child was determined and visited. The software was used to keep track of disease status and the total distance travelled. "Visiting" the nearest household containing a selected child was repeated until all selected children were included. Since the sampling unit for SRS and stratified random sampling was the eligible child, only households with eligible children were "visited".

For EPI cluster sampling, the sampling unit was the household. A household was selected at random from a list of all households falling within a path of a defined width from the ward centre in a randomly chosen direction. The house was then examined to determine if a child of eligible age was living there. Subsequently, the nearest (or *n*th nearest) household to the right was "visited", and the steps repeated until the desired number of children was obtained. The prevalence of diarrhoea and dysentery and the total distance travelled, including the distance from the centre of the ward to the starting household, were determined. The total distance travelled also included "visiting" households with no eligible children.

The "true" disease prevalence was considered to be the number of children with diarrhoea (or dysentery) as determined by the census, divided by the total number of enrolled children. Disease prevalence estimates from stratified random sampling and EPI cluster sampling were weighted according to the number of eligible children in each ward, in order to obtain estimates that could be compared with "true" prevalence. In a survey in which the primary sampling units (wards) would have been PPS selected, an unweighted prevalence estimate for the stratified random sample or the EPI cluster sample would have been appropriate because of the fixed sample size in each cluster. A prevalence estimate was obtained for each of the 1000 simulations run for each method. The mean prevalence of the 1000 simulations was then compared with true prevalence to estimate the bias of each method. Variance was obtained from the distribution of the 1000 prevalence estimates.

The effect of sampling design (design effect) on the variance of estimates obtained with EPI cluster sampling was calculated by dividing the estimate variance for EPI cluster sampling by the estimate variances for SRS and stratified random sampling. Thus, two different design effects were calculated. One answers the question "How much does estimate variance increase if cluster sampling instead of SRS is used to assess the prevalence of diarrhoea and dysentery?" the second answers the question "For the wards selected using PPS sampling at the first stage, how much does estimate variance increase using EPI cluster sampling instead of SRS for the second sampling stage?" Mean square errors (MSE) were calculated from the bias and variance estimates derived from stratified random and EPI cluster sampling by adding the square of bias to variance. MSE represents the overall error of prevalence estimates under different sampling designs.

Estimates of distance travelled were compared for EPI cluster sampling and stratified random sampling for each simulation run. Distance was compared ward by ward because ward maps were not drawn to a common scale and mapping distortion was variable. Median distance was calculated for each ward for each sampling method. A distance ratio was calculated for each ward by dividing the median distance travelled for EPI cluster sampling

by median distance travelled for stratified random sampling. A ratio greater than 1 indicates that EPI cluster sampling required a greater travel distance.

Maps of households visited for EPI cluster sampling (Fig. 1) and stratified random sampling (Fig. 2) are shown for ward 1.2. In Fig. 1, of the eight houses visited during EPI cluster sampling simulation, eligible children were found in only three. The distance travelled is greater for children selected at random (Fig. 2) than for those selected using the EPI sampling method (Fig. 1).

MSE measures the total error of sample estimates and distance can be used as an indicator of cost. To compare two designs (a & b), one can use the efficiency ratio (*20*):

$$\frac{Cost_a \times MSE_a}{Cost_b \times MSE_b}$$

By these criteria, a ratio greater than 1 implies design b is preferred, i.e. it has either a smaller cost per unit MSE or a smaller MSE per unit cost. The product of the cost and MSE of estimates obtained from EPI cluster sampling was compared to that of stratified random sampling for the same sample size.

Limitations

Distance travelled can provide only an approximate indication of the cost of conducting a survey. Other costs may include personnel, training, and materials. However, for comparison across sample designs, distance travelled should provide reasonably accurate estimates. However, if the cost of obtaining an accurate sampling frame is considered, SRS may be far more costly.

Distance was calculated as a straight line between two points. Roads, rivers, and other physical obstacles were not taken into consideration. Again, however, all simulations were subject to the same constraints, so comparison of designs should not have been seriously affected.

Finally, a field worker cannot be perfectly modelled using a computer. For example, the role of individual subjectivity in household selection cannot be simulated. As the selection algorithm used in the simulations is completely objective and consistent, the simulations represent an "ideal" that may be more or less close to the real situation.

Results

A total of 4297 (90.2%) of 4765 eligible children lived in 54.7% of the 4920 households in the 40 wards under study. The number of eligible children in each ward ranged from 13 to 284, whereas the number of households ranged from 31 to 315. The "true" prevalence of diarrhoea was 13.8%, and the prevalence of dysentery 4.4%, among the 4297 children surveyed.

Fig. 1 **Map of ward 1.2 showing the households visited (starting with household 27) for EPI cluster sampling** (sample size 7, next nearest household to the right visited)

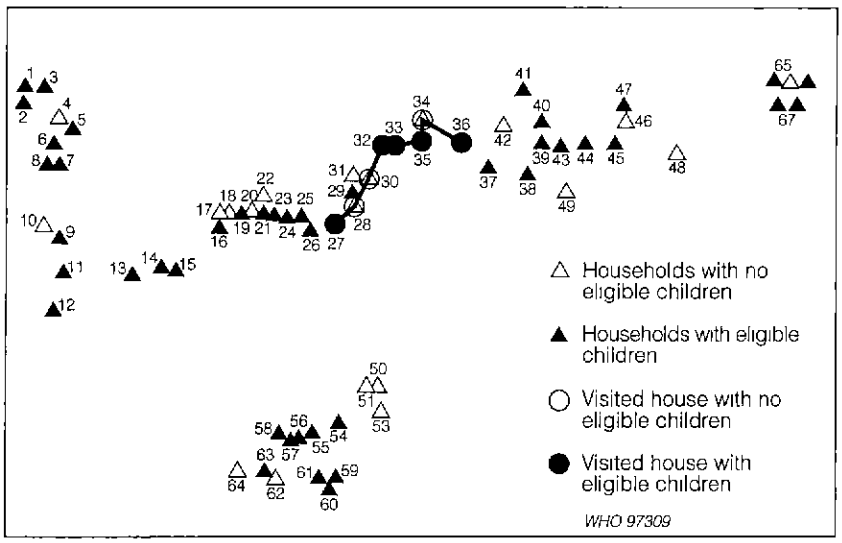
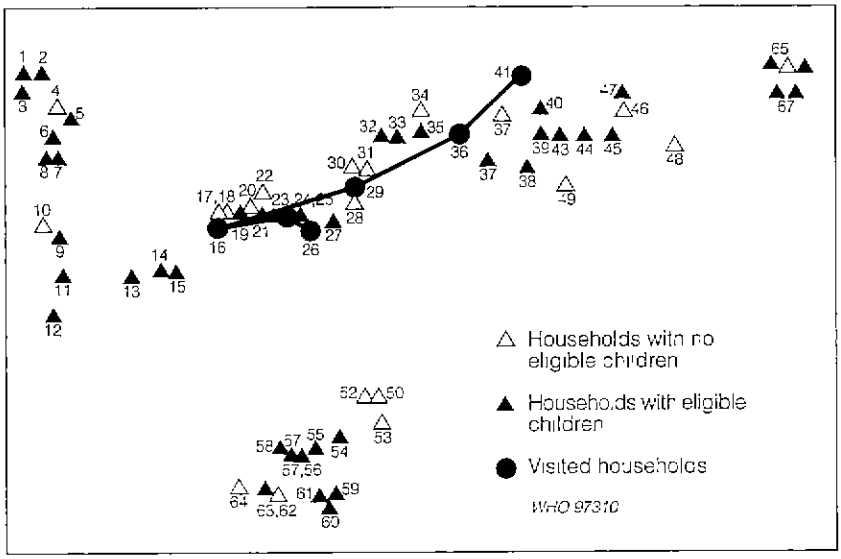


Fig 2. **Map of ward 1.2 showing households visited (starting with household 26) for stratified random sampling** (sample size 7).



Estimates based on simulations ranged from 13.75% (simple random sampling, sample size 10) to 16.45% (EPI cluster sampling, sample size 7, 4th nearest house to the right selected) for diarrhoea (Table 1) and 3.85% (EPI cluster sampling, cluster size 25, nearest house to the right selected) to 5.4% (EPI cluster sampling, cluster size 7, 3rd nearest house to the right selected) for dysentery (Table 2). For diarrhoea, EPI cluster sampling always resulted in higher prevalence than "true" prevalence.

Bias, calculated as the difference between true and sample prevalence, was very small for stratified random sampling. However, bias from EPI cluster sampling ranged from 0.36% to 2.65% (diarrhoea) and −0.55% to 1.00% (dysentery) (Fig. 3 & 4). EPI-derived estimates for diarrhoea prevalence showed positive bias (EPI cluster sampling consistently overestimated diarrhoea prevalence). Bias for EPI-derived dysentery prevalence estimates was either positive or negative, depending on the number of houses to the right that were skipped. The typical EPI procedure of selecting the next nearest household consistently underestimated the prevalence of dysentery (negative bias) regardless of sample size.

Estimate variances for EPI cluster sampling were compared to estimate variances for SRS to estimate the design effect of EPI cluster sampling. This estimates the overall increase or decrease in estimate

Table 1: **Prevalence estimates for diarrhoea obtained from EPI cluster sampling and simple random sampling (SRS)**

Sample size	True prevalence	EPI (1)[a]		EPI (2)		EPI (3)		EPI (4)		EPI (5)		SRS	
		Mean	SD[b]	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
7	13.80	15.76	2.31	16.00	2.34	15.85	2.47	16.45	2.28	15.58	2.09	13.80	2.53
10	13.80	15.79	1.77	15.68	1.95	16.11	1.98	15.46	1.76	15.32	1.78	13.75	2.06
15	13.80	15.06	1.32	14.95	1.49	15.91	1.37	14.63	1.23	14.35	1.24	13.78	1.65
20	13.80	14.94	1.08	14.71	1.25	15.27	1.04	14.23	1.00	14.16	1.02	13.82	1.42
25	13.80	15.35	0.80	14.28	0.99	15.01	0.80	14.57	0.89	14.27	0.86	13.81	1.23

[a] Figures in parentheses are the number of households to the right that were skipped.
[b] SD = standard deviation.

Table 2: **Prevalence estimates for dysentery obtained from EPI cluster sampling and simple random sampling (SRS)**

Sample size	True prevalence	EPI (1)[a]		EPI (2)		EPI (3)		EPI (4)		EPI (5)		SRS	
		Mean	SD[b]	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
7	4.40	4 02	1 26	5 07	1.58	5.40	1 49	4 82	1 39	4 99	1.46	4.44	1.49
10	4 40	4 05	1 00	4.81	1.32	5.13	1 17	4 47	1.04	4.58	1.08	4.45	1.23
15	4 40	4 09	0.76	4.78	0.88	4 71	0 75	4.02	0.73	4.17	0.73	4 41	0.98
20	4.40	3.91	0.63	4.49	0 66	4 49	0 62	4.06	0.58	4.02	0 61	4 40	0 82
25	4.40	3.85	0.50	4.13	0 55	4 30	0.49	3.90	0.44	3.97	0 55	4 42	0 72

[a] Figures in parentheses are the number of the households to the right that were skipped.
[b] SD = standard deviation

variance due to use of the EPI cluster sampling design compared with the random selection of children from the entire census list (Table 3). The design effect for estimates of diarrhoea prevalence ranged from 0.47 to 1.52 for EPI cluster sampling versus SRS. The value 0.47 was obtained from EPI cluster sampling with a sample size of 25 when the nearest house to the right was visited. The value 1.52 was obtained from EPI cluster sampling with a sample size of 7 when the 3rd nearest house to the right was visited. For estimates of dysentery prevalence, the design effect ranged from 0.46 (sample size 25, 4th nearest house to the right) to 1.76 (sample size 7, 2nd nearest house to the right) (Table 3). In general, larger sample sizes resulted in smaller design effects. The number of houses to the right that were skipped did not influence the design effect.

The comparison of estimate variances from EPI cluster sampling with those from stratified random sampling estimates the impact of using EPI cluster sampling instead of random sampling at the second sampling stage. This gives an estimate of that portion of the overall increase or decrease in variance associated with EPI cluster sampling that is due solely to the second sampling stage. The design effect for estimates of diarrhoea prevalence ranged from 0.42 (sample size 25, nearest house or 3rd nearest house to the right) to 0.95 (sample size 7, 3rd nearest house to the right) (Table 3). For estimates of dysentery prevalence, the design effect ranged from 0.37 (sample size 25, 4th nearest house to the right) to 1.15 (sample size 10, 2nd nearest house to the right) (Table 3). These design effects from EPI cluster sampling versus stratified random sampling were

Fig 3. **Prevalence estimates for diarrhoea, by number of households to the right skipped and sample size.** (SRS = simple random sampling, EPI(n) = n households to the right skipped)

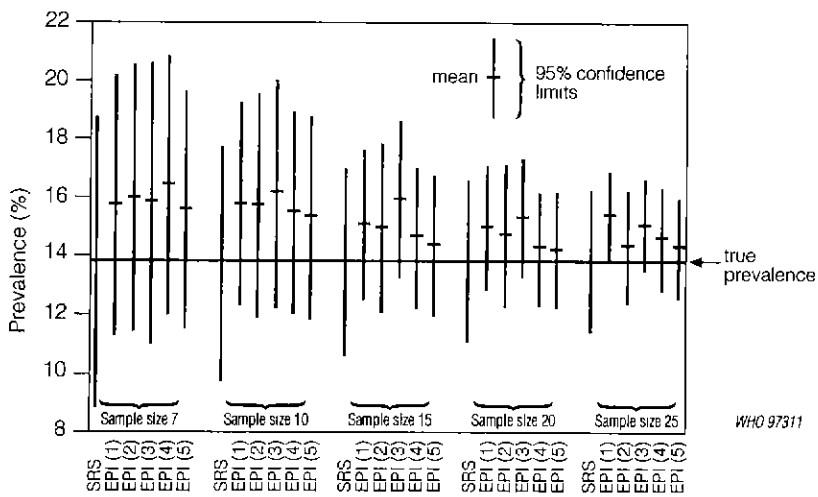


Fig. 4. **Prevalence estimates for dysentery, by number of households to the right skipped and sample size.** (SRS = simple random sampling, EPI(n) = n households to the right skipped)

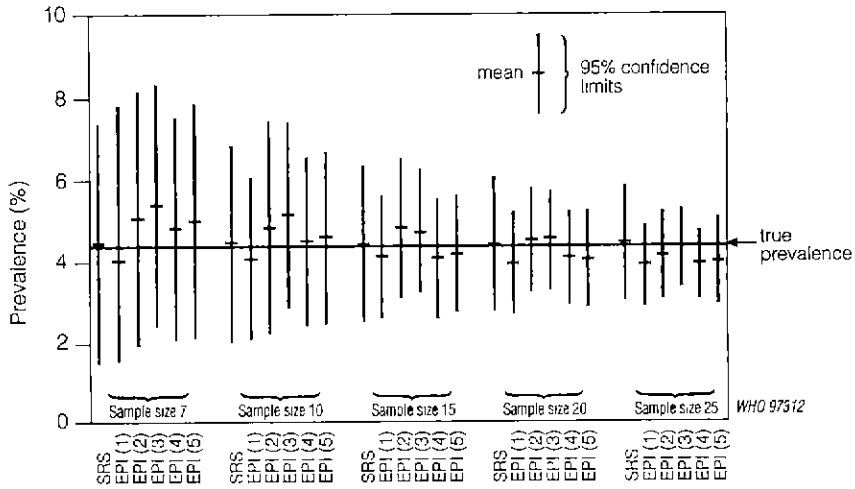


Table 3: **The design effects of EPI cluster sampling**

Design	Diarrhoea[a]	Diarrhoea[b]	Dysentery[a]	Dysentery[b]
EPI (1, 7)[c]	1.33	0.83	1.12	0.72
EPI (1, 10)	1.01	0.73	0.98	0.65
EPI (1, 15)	1.02	0.64	0.90	0.59
EPI (1, 20)	0.98	0.58	0.95	0.60
EPI (1, 25)	0.47	0.42	0.61	0.48
EPI (2, 7)	1.36	0.85	1.76	1.13
EPI (2, 10)	1.25	0.91	1.73	1.15
EPI (2, 15)	1.30	0.81	1.22	0.80
EPI (2, 20)	1.31	0.77	1.05	0.66
EPI (2, 25)	0.72	0.65	0.73	0.58
EPI (3, 7)	1.52	0.95	1.56	1.00
EPI (3, 10)	1.27	0.92	1.35	0.89
EPI (3, 15)	1.10	0.69	0.89	0.58
EPI (3, 20)	0.92	0.54	0.90	0.57
EPI (3, 25)	0.47	0.42	0.59	0.46
EPI (4, 7)	1.30	0.81	1.35	0.87
EPI (4, 10)	1.00	0.73	1.07	0.71
EPI (4, 15)	0.88	0.55	0.86	0.56
EPI (4, 20)	0.84	0.50	0.81	0.51
EPI (4, 25)	0.57	0.52	0.46	0.37
EPI (5, 7)	1.09	0.68	1.50	0.96
EPI (5, 10)	1.03	0.75	1.16	0.77
EPI (5, 15)	0.90	0.56	0.86	0.56
EPI (5, 20)	0.87	0.51	0.88	0.55
EPI (5, 25)	0.54	0.49	0.73	0.58

[a] Compared with simple random sampling
[b] Compared with stratified random sampling
[c] Figures in parentheses are, respectively, the number of households to the right that were skipped and the sample size

all less than one, except for those associated with estimates of dysentery prevalence with sample sizes 7 or 10, (2nd nearest house to the right).

The ratio of median distances (cost) increased when either n increased (nth nearest house to the right) or sample size increased (Table 4). The median distance ratios were all less than one when either nearest neighbours or next nearest neighbours were visited, regardless of sample size. This meant that for EPI cluster sampling visiting either the nearest house or the second nearest house, the distance travelled was always less than distance travelled for stratified random sampling to obtain the same number of subjects. If either the 3rd, 4th, or 5th nearest household was visited, the distance ratio was greater than one for large sample sizes. In other words, by increasing n and cluster size, the distance travelled using EPI cluster sampling increased more than the distance travelled for similar sample sizes using stratified random sampling. The cost benefit of EPI cluster sampling (at least for small neighbourhoods within a village) is negated when households farther away than the 2nd nearest to the right are visited.

Efficiency ratios based on comparisons of the MSE and cost of prevalence estimates for diarrhoea obtained from EPI cluster sampling versus stratified random sampling increased with the sample size

Table 4: **Median distance ratios of EPI cluster sampling compared with stratified random sampling**

	*n*th nearest household to the right sampled.				
Cluster size	1	2	3	4	5
7	0 39	0 53	0.64	0 72	0 82
10	0 42	0 58	0.68	0 80	0 92
15	0.49	0 62	0 80	1 01	1 07
20	0 55	0.77	0 92	1.28	1 33
25	0 65	0.95	1 13	1.47	1 62

when the nearest household to the right was visited (Table 5). The values ranged from 0.58 (sample size 7, nearest household to the right) to 1.97 (sample size 15 3rd nearest household to the right). Efficiency ratios compare EPI cluster sampling and stratified random sampling, and an efficiency ratio less than one indicates that EPI cluster sampling is more efficient than stratified random sampling, considering both distance travelled and size of MSE. Similar results were obtained for estimates of dysentery prevalence: ratios ranged from 0.31 (sample size 7, nearest household to the right) to 1.51 (sample size 25, 5th nearest household to the right). The ranking

Table 5: **The ratios of the products of distance travelled and mean square error (efficiency ratios) for EPI cluster sampling versus stratified random sampling**

Design	Diarrhoea	Dysentery
EPI (1, 7)[a]	0 58	0.31
EPI (1, 10)	0 74	0 31
EPI (1, 15)	0 64	0.34
EPI (1, 20)	0.73	0 53
EPI (1, 25)	1 44	0 69
EPI (2, 7)	0 89	0 71
EPI (2, 10)	1 06	0 73
EPI (2, 15)	0 85	0.59
EPI (2, 20)	0 98	0.51
EPI (2, 25)	0 82	0.68
EPI (3, 7)	1 07	0.93
EPI (3, 10)	1 56	0.85
EPI (3, 15)	1 97	0.55
EPI (3, 20)	1 61	0 53
EPI (3, 25)	1 75	0 54
EPI (4, 7)	1 44	0.68
EPI (4, 10)	1 16	0.57
EPI (4, 15)	0 87	0.72
EPI (4, 20)	0.81	0 87
EPI (4 25)	1.48	1 24
EPI (5, 7)	1 01	0.92
EPI (5, 10)	1.25	0.73
EPI (5, 15)	0 76	0 66
EPI (5, 20)	0 82	1.02
EPI (5, 25)	1 13	1.51

[a] Figures in parentheses are, respectively, the number of households to the right that were skipped and the sample size

of efficiency ratios on the basis of the nearness of households visited for a given cluster size was different for diarrhoea and dysentery, suggesting that the two diseases have different patterns of spatial distribution.

Discussion

EPI cluster sampling is often used when resources are limited, as it is easier and quicker to carry out than SRS or stratified random sampling. However, EPI cluster sampling can result in increased estimate bias and variance compared with SRS or stratified random sampling. In this study, EPI cluster sampling resulted in prevalence estimates for diarrhoea that were consistently larger than the prevalence as determined by census, regardless of sample size or the number of houses to the right that were skipped. Estimates for the prevalence of dysentery varied. Some showed positive bias and others negative, depending on sample size and the number of households skipped (Tables 1 & 2).

A possible explanation for the different bias patterns is that diarrhoeal disease may tend to be more concentrated in the centre of a village; as EPI cluster sampling tends to select starting households that are central rather than peripheral, diarrhoea prevalence may be overestimated. EPI cluster sampling may therefore introduce bias in prevalence estimates according to the actual pattern of disease distribution Diarrhoea and dysentery were examined in this study because they are infectious diseases and may be transmitted by a single source of infection In addition, children from the same neighbourhood often play together, which may add to infectious disease transmission and shape the resulting spatial distribution of disease

On the basis of this study, we conclude that diarrhoea prevalence estimated with EPI cluster sampling would be an overestimate of the prevalence of diarrhoea. If the accuracy of a single survey is important, EPI cluster sampling may not be the right tool. However, if surveys are used to assess changes in prevalence over a period of time, EPI cluster sampling may be appropriate. EPI cluster sampling would in any case be less expensive and quicker to carry out than SRS or stratified random sampling.

For less common diseases such as dysentery, EPI cluster sampling may be appropriate for assessing community disease prevalence when either accuracy or the measurement of trends is important. No systematic bias due to EPI cluster sampling was found when estimating dysentery prevalence, and there was very little increase in variance associated

with EPI cluster sampling compared with stratified random sampling. However spatial clustering can occur even with diseases of low prevalence, and EPI cluster sampling may in such a case produce biased estimates.

The spreading of sampled households over a larger geographical area by increasing n when selecting every nth household to the right should theoretically result in more heterogeneous samples and greater precision of estimates. It may also reduce their bias, since sampling from a larger area may minimize the influence of the "neighbourhood" effect (i.e., similarities in variables of interest between houses located near each other). Another way to obtain more heterogeneous samples is to interview only one child per household. This has the benefit of removing within-household similarities from the sample (e.g. common morbidities among siblings). In this population the effects of within-household similarities were larger than within-village similarities (*21*). In our study, neither of these modifications affected the variance or the bias.

Although both diarrhoea and dysentery prevalence estimates showed substantial relative bias (diarrhoea, up to 17%; dysentery up to 23%), the bias for estimates of diarrhoea was always positive, whereas the bias for estimates of dysentery varied. By increasing the sample size, variance decreased, and, to a lesser extent, so did bias for estimates of both diarrhoea and dysentery prevalence. This can be explained by the increasing proportion of the total population sampled. The largest sample size (1000) included approximately 1 out of every 4 children in the population. In some small wards, when the largest sample size was used, every child was selected.

Further studies would be needed to determine whether these results can be applied in other settings, or whether seasonal variations in disease prevalence would alter these findings. Modifications to EPI cluster sampling may provide more accurate estimates and should be examined more closely, perhaps using computer simulation techniques (*22*).

Acknowledgements

This work was carried out under Cooperative Agreement No. DAN 0045-A-5094 (Office of Health and Nutrition, United States Agency for International Development (USAID), Washington, DC & Center for Human Nutrition and Dana Center for Preventive Ophthalmology (CHN/DCPO), Johns Hopkins University, Baltimore, MD), with additional support from National Institute of Health (NIH) grant No. RR04060. The authors wish to thank Dr S. Bennett of the London School of Hygiene and Tropical Medicine for his helpful advice, and colleagues at the Nepal Nutrition Intervention Project, Sarlahi, Nepal (Dr S.K. Khatry, S. LeClerq, E. Pradhan, S.R. Shrestha, N.N. Acharya, D.N. Mandal, T.R. Sakiya) and the National Society for the Prevention of Blindness (Dr R.P. Pokhrel), Kathmandu, Nepal.

Résumé

Efficience du sondage en grappes du PEV pour l'estimation de la prévalence de la diarrhée et de la dysenterie

La diarrhée et la dysenterie sont des causes majeures de morbidité et de mortalité chez l'enfant dans de nombreux pays en développement; le traitement et la prévention ne sont toutefois possibles que si l'on dispose en temps voulu d'informations exactes sur leur prévalence. Un système de surveillance utilisant des sondages en cours permet d'obtenir de telles données.

Plusieurs méthodes d'enquête par sondage sont utilisables, mais les ressources disponibles et le niveau d'exactitude requis sont des facteurs contraignants. Le sondage aléatoire simple, s'il est fait correctement, fournit des estimations précises et exactes (c'est-à-dire des estimations non biaisées, ayant une variance faible). Les conditions de réalisation du sondage aléatoire simple sont toutefois difficiles, voire impossibles, à remplir. Il est en effet particulièrement malaisé d'obtenir une base de sondage appropriée (c'est-à-dire une liste de tous les membres de la population), de plus, si la population est géographiquement dispersée, et que certains membres habitent des secteurs éloignés ou inaccessibles, les entretiens risquent de poser des problèmes logistiques supplémentaires.

Une méthode simplifiée de sondage en grappes mise au point par le Programme élargi de Vaccination (PEV) de l'OMS permet de résoudre ces difficultés, en utilisant à la fois une base de sondage en grappes, et une stratégie de sondage qui limite la dispersion géographique des sujets sélectionnés. Ces avantages sont toutefois obtenus au prix de la perte d'une certaine exactitude et d'une certaine précision des estimations obtenues. Cette étude se propose de mesurer comment l'économie de ressources compense l'augmentation du biais et de la variance des estimations.

On constate que la méthode de sondage en grappes du PEV présente systématiquement un biais relatif positif qui atteint 17% (par rapport à la moyenne pour la population) pour les estimations de prévalence de la diarrhée. Concernant la dysenterie, le biais est tantôt positif, tantôt négatif. Les estimations de prévalence de la diarrhée sont plus précises que celles obtenues avec le sondage

en grappes classique (sondage aléatoire stratifié); les estimations de prévalence de la dysenterie sont tantôt plus précises, tantôt moins précises, que celles obtenues par sondage en grappes classique.

Le sondage en grappes du PEV a exigé en général des déplacements moins lointains (la distance sert d'indicateur de coût dans l'analyse) que le sondage en grappes classique; toutefois, comme on s'y attendait, cet avantage a disparu au fur et à mesure que des modifications destinées à garantir une plus grande dispersion géographique des sujets étaient apportées à la méthode du PEV.

Dans la mesure où, avec le sondage en grappes du PEV, l'on observe une variation du biais avec la maladie étudiée, il est possible que les différences de répartition spatiale des variables considérées (morbidité par exemple) modifient les estimations de prévalence lorsqu'on utilise cette méthode. Les estimations obtenues grâce à la méthode du PEV peuvent donc avoir une exactitude acceptable dans certaines situations. Provisoirement, ces estimations seront toutefois interprétées avec une certaine prudence. Reste que le biais étant constant, comme c'est le cas pour les estimations de prévalence de la diarrhée, la surveillance des tendances dans la population n'en est pas affectée.

References

1. **Sachdev HPS et al.** Risk factors for fatal diarrhea in hospitalized children in India. *Journal of pediatric gastroenterology and nutrition*, 1991, **12**: 76–81.
2. **Savarino SJ, Bourgeois AL.** Diarrhoeal disease current concepts and future challenges Epidemiology of diarrheal diseases in developed countries *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 1993, **87** (suppl. 3): 7–11
3. **Black RE, Brown KH, Becker S.** Malnutrition is a determining factor in diarrhea duration, but not incidence, among young children in a longitudinal study in rural Bangladesh *American journal of clinical nutrition*, 1984, **37**: 87–94.
4. **Pelletier DL et al.** The effects of malnutrition on child mortality in developing countries. *Bulletin of the World Health Organization*, 1995, **73** 443–448
5. **Black RE et al.** Contamination of weaning foods and transmission of enterotoxigenic *Escherichia coli* diarrhoea in children in rural Bangladesh *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 1982, **76** 259–264.
6. **Cochran WG.** *Sampling techniques*, 3rd ed New York, Wiley, 1978: 233.
7. *Information system.* Geneva, World Health Organization, 1992 (unpublished document WHO/EPI/CEIS/92 1, available on request from the Expanded Programme on Immunization, World Health Organization, 1211 Geneva 27, Switzerland)
8. **Lemeshow S, Robinson D.** Surveys to measure programme coverage and impact: a review of the methodology used by the Expanded Programme on Immunization *World health statistics quarterly*, 1985, **38**. 65–75.
9. **Frerichs RR, Tar Tar K.** Computer-assisted rapid surveys in developing countries. *Public health reports*, 1989, **104**: 1, 14–23
10. **Dabis F et al.** Monitoring selective components of primary health care methodology and community assessment of vaccination, diarrhoea, and malaria practices in Conakry, Guinea. *Bulletin of the World Health Organization*, 1989, **67**: 675–684.
11. **Melgaard B, Mutie DM, Kimani G.** A cluster survey of mortality due to neonatal tetanus in Kenya. *International journal of epidemiology*, 1988, **17** 174–177
12. **Georges MC et al.** Diarrheal morbidity and mortality in children in the Central African Republic. *American journal of tropical medicine and hygiene*, 1987, **36**: 598–602.
13. **Auer C, Tanner M.** Childhood vaccination in a squatter area of Manila: coverage and providers *Social science and medicine*, 1990. **31**: 1265–1270
14. **Zollner E, Carlier ND.** Breast-feeding and weaning practices in Venda. 1990. *South African medical journal*, 1993, **83**: 580–583.
15. **West KP et al.** Efficacy of vitamin A in reducing preschool child mortality in Nepal, *Lancet*. 1991, **338**: 69–70.
16. **West KP et al.** Tolerance of young infants to a single, large dose of vitamin A a randomized community trial in Nepal. *Bulletin of the World Health Organization*, 1992, **70**: 733–739.
17. **Pradhan EK et al.** Data management for large community trials in Nepal, *Controlled clinical trials*, 1994. **15**: 220–234.
18. **Bennett S et al.** A computer simulation of household sampling schemes for health surveys in developing countries. *International journal of epidemiology*, 1994, **23** 1282–1291.
19. **Lemeshow S et al.** A computer simulation of the EPI survey strategy *International journal of epidemiology*, 1985, **14**: 473–481.
20. **Kish** L. *Survey sampling*, New York, Wiley, 1965: 266.
21. **Katz J et al.** Estimation of design effects and diarrhea clustering within households and villages. *American journal of epidemiology*, 1993, **138**: 994–1006.
22. **Bennett S et al.** A simplified general method for cluster-sample surveys of health in developing countries *World health statistics quarterly*, 1991, **44**: 98–106