Bulletin of the World Health Organization, 63 (1), 93 – 99 (1985)

© World Health Organization 1985

Observations on the application of EPI cluster survey methods for estimating disease incidence

R. B. ROTHENBERG,[1] A. LOBANOV,[2] K. B. SINGH,[3] & G. STROH, JR[4]

The present study attempted to assess the incidence of target diseases of the Expanded Programme on Immunization (poliomyelitis, tetanus, measles, pertussis, neonatal tetanus, diphtheria), using cluster samples and a household interview form. The results suggest that this method can indeed serve to estimate the incidence of these diseases with reasonable precision and may also be used to demonstrate reduction in incidence for the more common diseases. Analysis of 37 surveys for poliomyelitis and neonatal tetanus in India revealed a relative uniformity in the design effect (i.e., the ratio of the variance for the cluster estimate to the variance for the binomial estimate) for diseases with low incidence and prevalence. Diseases with higher prevalence tend to have a larger design effect, which may be indicative of the epidemic and "clustered" nature of the disease. A large design effect, therefore, does not necessarily indicate a need for a larger sample size, particularly if precision is acceptable. There is no one single design that is ideal for all surveys of disease incidence and decisions must be made in the light of local conditions and available resources.

The evaluation of a disease control programme rests ultimately on the demonstration of a reduction in disease incidence. In many industrial countries, this can be provided through analysis of data collected through both routine and special reporting systems. In developing countries, greater reliance must be placed on low-cost methods that are largely independent of the health infrastructure. Sentinel systems, investigation of outbreaks, and special surveys fulfil these needs to a greater or lesser extent, depending on local circumstances.

Cluster survey techniques have been developed and used to determine vaccine coverage of a population (*1*). Similar methods, with appropriately expanded sample sizes, have been applied to the detection of poliomyelitis and neonatal tetanus (*2*).[a] In the present study, we applied a similar approach to assess the incidence of selected target diseases of the Expanded Programme on Immunization (EPI)—poliomyelitis, neonatal tetanus, tetanus, diphtheria, pertussis, and measles—in areas of Nepal.

[1] Director, Bureau of Chronic Disease Prevention, New York State Department of Health, Albany, NY 12237, USA

[2] Medical Officer, Expanded Programme on Immunization, World Health Organization, Geneva, Switzerland.

[3] Programme Manager, Expanded Programme on Immunization, Kathmandu, Nepal.

[4] Chief, Training and Development Branch, International Health Program Office, Centers for Disease Control, Atlanta, GA, USA.

[a] BERNIER, R. H. Some observations of polio lameness surveys. Presented at the *International Symposium on Worldwide Polio Eradication, Washington, DC, March 1983*.

METHODS

Two household cluster surveys were performed in Nepal between November 1982 and February 1983, in the urban and semiurban areas of the Kathmandu Valley and in the district of Dhanusa. The standard EPI method (*3*) entails first-stage sampling of defined populations in an area, with probability of inclusion proportional to size of population. The second stage is performed by random selection of the first household and completion of the sample with neighbouring households. The survey teams used a standard form (see Annex) on which they recorded information concerning household size, the number of living children under 1 year, the number of children aged between 1 and 9 years, the number aged 5–9 years, and the number of children in each of these categories who had died in the previous 12 months.

The respondent was also asked about disease occurring during the preceding 12 months. A short standard description, together with illustrative material, was used to identify the six target diseases (Table 1). Information was sought on each child, living or dead, appropriate to his or her age category—neonatal tetanus, 3–28 days after birth; poliomyelitis, 5–9 years; tetanus, pertussis, diphtheria, and measles, all age categories. For poliomyelitis and neonatal tetanus, case verification was attempted using the case investigation form developed for the neonatal tetanus/poliomyelitis surveys performed in India (*3*). For the other

Table 1. Definitions for target diseases used in household survey

Symptoms	Disease
10-day rash, illness with fever, cough, coryza and conjunctivitis	Measles
Severe sore throat with difficulty in breathing, fever and systemic toxicity	Diphtheria
Upper respiratory infection with non-productive cough and typical "whoop"	Pertussis
Tightness, spasticity, rigidity and paralysis of breathing	Tetanus
Failure to suck, weak cry, spasticity, seizures, and death in the first 3–28 days of life	Neonatal tetanus
Flaccid atrophy of leg muscle with limp in otherwise healthy child	Poliomyelitis

diseases, recognition by the parents of the typical syndrome was accepted as evidence of the disease.

The sample size for the survey was chosen so as to permit detection of neonatal tetanus with a precision of 50% (precision is defined here as the ratio of the 95% confidence limits to the estimate). With an expected incidence of 2% for neonatal tetanus, this generated a sample size of 800, using the standard formula for the variance of a binomial variable:[b]

$$V = (pq/n)\,((N-n)/N),$$

where p = probability of having had the disease,
$q = 1-p$,
n = sample size,
and N = population size.

The finite population correction (fpc) was assumed to be negligible.

The design effect, first described by Cornfield in 1951 (*4*), is defined as the ratio of the variance for the cluster estimate to the variance of the binomial estimate. A previous neonatal tetanus/poliomyelitis survey in Nepal[c] indicated that the design effect for a cluster sample would be about 1.5. It was decided not to consider the design effect and to accept the loss of precision that might occur.

The proposed sample of 800 infants under one year was divided into 40 clusters of 20 infants each. It was estimated from demographic data that an average of 2400 children aged 1–4 years and 3000 children aged 5–9 years would be found in the households visited to obtain the 800 infants.

Training of the survey teams and supervisory staff was performed over several days. A system for accumulating household data on a cluster summary form was used to ensure stepwise recording of data. The data were analysed for each disease separately, taking the overall estimate of disease incidence as $p = (\Sigma\, a_i)/(\Sigma\, m_i)$ where a_i is the number of children in the ith cluster who have had the disease in the previous 12 months, and m_i is the number of subjects in the appropriate age group in the cluster. (Note: This formula presumes that children in the target age range constitute a fixed proportion of the total population in each cluster. The extent to which this assumption affects the outcome is currently under investigation by one of the present authors (G.S.). Preliminary analysis indicates that variability among clusters in the proportion of target-age children does not grossly distort the results.)

The variance for the estimate of cluster sample proportions was calculated as

$$V = (1-(n/N))\,\Sigma\,(a_i - pm_i)^2\,/(n(n-1)\bar{m}^2)$$

where n is the number of clusters and $\bar{m}$ is the average cluster size.[d] The finite population correction was 0.9337 for the Kathmandu study and 0.9346 for Dhanusa.

A critical ratio for the difference of proportion (z-test) was calculated to assess the statistical significance of the difference in rates between the urban and rural surveys (*11*). Box plots were drawn using standard methods (*12*).

RESULTS

The data from the surveys in Kathmandu and Dhanusa are displayed in Tables 2 and 3, in order of diminishing design effect. (The original survey data and details of calculations will be supplied by the authors on request).

These results suggest attack rates of 6% (rural) and 12% (urban) per year for measles in children under 9 years and 1.5% and 4% respectively for pertussis. The differences between urban and rural rates in both cases are statistically significant ($P < 0.01$). Diphtheria, tetanus and poliomyelitis are rare in both settings, and the differences observed were not significant. Neonatal tetanus appeared to be more common in the rural area, as has been observed previously, but this difference was also not significant.

There was a direct relationship between the incidence of disease and the design effect, and an

[b] It should be noted that the second stage of sampling in the EPI cluster survey method does not strictly adhere to the rules governing use of standard formulae for calculating the variance of a binomial variable. However, these standard formulae have been routinely used in EPI survey methods, with a recognition of this departure.

[c] ROTHENBERG, R. B. *Sample surveys to estimate the annual incidence of neonatal tetanus and poliomyelitis in the kingdom of Nepal.* Assignment report, SEARO/EPI/43, 1982.

[d] See footnote *b*.

Table 2. Results from Kathmandu survey for target diseases of the EPI

Disease	No. per 1000 per year[a]	Standard error	Precision[b] (%)	Design effect
Measles	123.9	9.0	14	7.8
Pertussis	40.5	4.5	21	5 0
Diphtheria	2 1	0.55	51	1.9
Neonatal tetanus	7.0	3.3	92	1 7
Poliomyelitis	1.3	0.45	67	1.18
Tetanus	0.0			

[a] Per 1000 individuals in the appropriate age group (see text).
[b] 95% confidence limits expressed as a percentage of the estimated rate; the smaller the percentage, the greater the precision.

Table 3. Results from Dhanusa survey for target diseases of the EPI

Disease	No. per 1000 per year[a]	Standard error	Precision[b] (%)	Design effect
Measles	56.6	11.8	41	16.9
Pertussis	14.9	4.5	59	8.8
Diphtheria	0.5	0 3	118	1.6
Neonatal tetanus	18.6	4.0	42	1.0
Poliomyelitis	2.6	1 1	83	1.0
Tetanus	0.2	0.19	186	0.8

[a] Per 1000 individuals in the appropriate age group (see text).
[b] 95% confidence limits expressed as a percentage of the estimated rate, the smaller the percentage, the greater the precision.

inverse relationship between these two factors and precision. The common diseases (measles and pertussis) exhibited reasonable precision under the conditions of this survey but had a large design effect, which presumably reflects focal occurrence. On the other hand, the design effect for the rarer diseases was small, suggesting random dispersion. Much larger samples would be required, whether by cluster or simple random sampling, to improve the precision of these estimates.

These observations are consistent with the theoretical expectation of a direct relationship between prevalence and design effect and an inverse correlation between these factors and precision. Table 4 presents the precision and design effect for diseases with estimated incidence from 0.001 to 0.20. In each instance, maximal clustering is assumed, in order to simulate the condition of maximal variance. It is immediately evident that common diseases that are markedly clustered—as would be the case with epidemics in isolated areas or chronic, scattered endemic foci—produce a large design effect, whereas a rare disease cannot cluster with the same statistical force. (For example, one case in a particular cluster and none in all the others, under the conditions assumed for Table 4, produces a sum of squared deviations from the mean of 0.975. Twenty cases in one cluster and none in the others generate a sum of squares of 399.2. Thus, the more common disease

Table 4. Theoretical precision and design effect for 40 clusters of 25 individuals each, with maximum aggregation of attributes ("worst case")

Estimated incidence	Standard error (cluster)	Precision (%)	Design effect
0.001	0 001	196	1.1
0.005	0.00487	190	4.1
0.01	0 00975	190	10.1
0.05	0.03402	133	24.4
0.10	0.04682	92	24.3
0.15	0.05573	73	25.9
0.20	0.06243	61	24.4

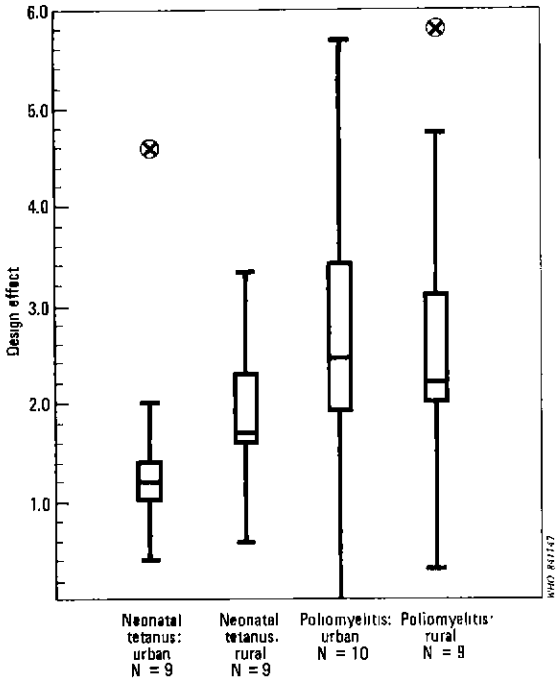

Fig. 1. Distribution of design effects for poliomyelitis and neonatal tetanus surveys in India. Box indicates fourth spread, crossbar shows median; extensions show outlier cutoff points; X indicates outlier. Compiled from data published by the Government of India and used by permission.

with the same maximal clustering produces a variance that is 400 times greater.)

Data from 10 surveys carried out in India on the incidence of poliomyelitis and neonatal tetanus in rural and urban areas (*5–10*)[e-h] provided a total of 37 estimates of design effect, which are displayed graphically in Fig. 1. (These data, together with their epidemiological interpretation and analysis of their effect on immunization programmes, will be the subject of a forthcoming communication by R. N. Basu et al., who have graciously permitted reference to them here.) It is observed that design effects for poliomyelitis were in general higher than those for neonatal tetanus and had a greater range. There was considerable overlap, however, with median values between 1.2 and 2.25. Only two outliers were recognized, one in the rural poliomyelitis surveys and one in the urban neonatal tetanus surveys. The data reported here for Nepal for the rare diseases conform to these distributions, with design effects in general between 1 and 2. For all 37 design effects and the corresponding estimates of disease incidence, there is a correlation coefficient of $r = 0.10$ ($P > 0.05$), suggesting that at the low end of the range for design effect and incidence, there is no correlation.

DISCUSSION

The surveys in Nepal measured the incidence of EPI target diseases during one year, and the prevalence of lameness due to poliomyelitis at the time of the survey. The one-year period is approximate, since family attention to dates is not fastidious. In addition, an exact 12-month interval probably represents only 11 months of risk, suggesting that the figures presented (except for poliomyelitis) are a slight overestimate.

The estimated incidence of measles (5–12%) was considerably lower than that estimated in 1975 during a serosurvey for the prevalence of measles in Nepal (*13*). Brink & Nakano, in 1978 (*14*), noted that the prevalence of measles antibody in children aged 60–72 months varied according to terrain (64% in hill regions, 81% in the Terai (marshland)) and was substantially higher than the 57% prevalence they observed in Sri Lanka. Previous serosurveys, using different serological techniques, suggested that 75% immunity levels can be seen as early as 37 months of age in some areas (*15,16*). The differences observed in the present survey may be attributed in part to a possible absence of recent epidemic activity. The 5–12% figure for one year's incidence would correspond to a prevalence of 25–60% for the age group, assuming constant incidence.

The results should thus be taken as an estimate of the true disease incidence, which could be ascertained only by expensive serological and culture methods. These are beyond the resources of many developing countries. If performed in a routine and rigorous way, cluster survey results can be reproducible and can be used to compare disease trends over time. This applies particularly to the more common diseases (neonatal tetanus, measles, and pertussis). The incidence of the rarer diseases (poliomyelitis, tetanus, and diphtheria) should perhaps be evaluated initially, but the sample sizes required for precise estimates and

[e] ASLANIAN, R. G. *Sample survey on deaths from neonatal tetanus and on the prevalence of poliomyelitis in the states of Punjab and Haryana and the Union Territory of Chandigarh (India)*. Assignment report, SEARO/EPI/23, 1981.

[f] GOODMAN, R. A. *Expanded Programme on Immunization, Indian sample survey of neonatal tetanus and paralytic poliomyelitis in Kerala State*. Assignment report, SEARO/EPI/40, 1982.

[g] MARTCHENKO, G. P. *Results of sample survey of neonatal tetanus and poliomyelitis deaths in Madhya Pradesh*. Assignment report, SEARO/EPI/24, 1981.

[h] POLLACK, M.P. *Sample survey to estimate the annual incidence of neonatal tetanus deaths and poliomyelitis in India (Allahabad Division of Uttar Pradesh and Jaipur in Rajasthan)*. Assignment report, SEARO/EPI/24, 1981.

for follow-up comparisons to demonstrate a decrease in incidence, may be prohibitive. Fortunately, the more common diseases are good indicators of the impact of the EPI. A reduction in pertussis, for example, may be evidence for widespread and consistent use of diphtheria-pertussis-tetanus (DPT) vaccine. A decrease in neonatal tetanus reflects the efficacy of the tetanus toxoid programme for pregnant women.

The distribution of design effect has important epidemiological implications. The empirical data obtained from the surveys in India and Nepal substantiate the hypothesis that design effect is inversely related to disease incidence in the situation where the common diseases are not uniformly distributed. Under the conditions of these surveys, adequate precision is maintained for the common diseases, but the design effect in one instance was 17. Such a high design effect is usually viewed with alarm, since it represents the amount by which the sample size for a cluster survey would have to be multiplied to produce a variance equal to that for a random sample. However, if precision is acceptable, there is no need to increase sample size. Rather, the design effect may be viewed as a measure of the degree to which the disease is clustered, i.e., the degree to which it occurs in focal epidemic form or in chronic endemic foci.

The operational feasibility and consistency of results of vaccine coverage surveys, using 30 clusters of 7 children each, rest on their simplicity, their ease of performance in the field, the immediate comprehensibility of their results and their relatively low cost. A disease incidence survey is in general a larger undertaking, and the size and logistics must be tailored to fit the resources and circumstances of the country. In India, 30 clusters of 67 infants aged 1–4 months, and all children aged 5–9 years encountered in the households visited to find the 67 infants, were used to study the prevalence of poliomyelitis and mortality due to neonatal tetanus. These surveys were carried out in both urban and rural areas. In Nepal, a single survey was performed for the entire country, using a total of 64 clusters containing 30 infants each, on average, and distributed among four districts. The problems involved in travel to remote mountainous areas, and the scarcity of resources, necessitated this approach.

Similarly, the use of 40 clusters of 20 infants each for the six-disease survey was determined by local constraints on transportation, manpower, and time. In these three situations, the design effects for the poliomyelitis and neonatal tetanus surveys were similar (usually between 1 and 3) indicating that considerable flexibility is available to the programme manager in choice of sample size, as well as in the choice of number, size, and distribution of clusters. This suggests that a single standard design, similar to the one for vaccine coverage, may not be ideal for surveys of disease incidence. The use of 30 clusters of 20–30 infants gears the study to the ascertainment of neonatal tetanus, and is offered as a starting point in planning the survey. A sample size of 800 or more infants would ensure reasonable precision for measles and pertussis as well, and provide baseline information for detection of any decrease and measurement of programme impact on these two diseases. The EPI manager must, however, pick a sampling scheme that fits the circumstances of the country.

As a final cautionary note, it should be mentioned that a disease incidence survey should be conducted in the context of the long-term EPI plan. Judicious site selection, programme implementation over several years, and careful timing of a follow-up survey, are important to minimize cost and maximize the likelihood of demonstrating an effect. The cluster survey method, under such circumstances, may prove to be a useful tool for documenting disease containment.

ACKNOWLEDGEMENTS

The authors acknowledge the help of Dr Roger Bernier of the Centers for Disease Control, and Dr Rafi Aslanian of the WHO Regional Office for South-East Asia EPI, who contributed to the design, implementation, and analysis of this work.

RÉSUMÉ

UTILISATION DES MÉTHODES DE SONDAGE PAR GRAPPES DU PEV POUR ESTIMER L'INCIDENCE D'UNE MALADIE: OBSERVATIONS

Les enquêtes visant à déterminer le niveau de couverture vaccinale obtenu grâce au programme élargi de vaccination (PEV) ont prouvé l'utilité de la méthode de sondage par grappes. Cette méthode a également été appliquée pour

mesurer la prévalence de la poliomyélite et la mortalité due au tétanos du nouveau-né. La présente étude avait pour but d'évaluer l'incidence des maladies cibles du PEV (poliomyélite, tétanos, rougeole, coqueluche, diphtérie) en procédant à un sondage par grappes et en utilisant un questionnaire adressé aux ménages. Les résultats indiquent que cette méthode permet effectivement d'estimer l'incidence de ces maladies avec un degré de précision raisonnable et de faire ressortir la réduction obtenue pour les maladies les plus courantes. L'analyse de 37 enquêtes sur la poliomyélite et le tétanos du nouveau-né en Inde a montré que le biais d'échantillonnage (c'est-à-dire le rapport des variances correspondant l'une à l'estimation par la méthode des grappes, l'autre à la variable binomiale) est relativement uniforme pour les maladies dont l'incidence et la prévalence sont faibles. Pour les maladies de prévalence élevée, le biais d'échantillonnage a tendance à être supérieur, ce qui pourrait être révélateur de la nature épidémique et de la distribution par grappes de la maladie. Un biais d'échantillonnage important ne signifie donc pas nécessairement qu'il faille choisir un échantillon plus grand, surtout si le degré de précision est suffisant. Il n'y a pas de mode d'échantillonnage qui convienne à toutes les enquêtes sur l'incidence, aussi la décision doit-elle être prise en fonction des conditions locales et des ressources disponibles sur place.

REFERENCES

1. HENDERSON, R. H. & SUNDARESAN, T. Cluster sampling to assess immunization coverage—a review of experience with a simplified sampling method. *Bulletin of the World Health Organization*, **60**: 253–260 (1982).
2. BASU, R. N. Magnitude of problem of poliomyelitis in India. *Indian pediatrics*, **18**: 507–511 (1981).
3. Combined survey to estimate the incidence of neonatal tetanus and poliomyelitis. *EPI bulletin (India)*, **14**: 24–33 (1981).
4. CORNFIELD, J. Modern methods in the sampling of human populations. *American journal of public health*, **41**: 654–661 (1951).
5. *Sample survey to estimate the incidence of poliomyelitis and neonatal tetanus in Delhi*. Delhi, Epidemiology Section, Municipal Corporation, 1981.
6. *Sample survey to estimate the incidence of poliomyelitis and neonatal tetanus in West Bengal*. Calcutta, EPI Section, Directorate of Health Service, Government of West Bengal, 1981.
7. *Sample survey to estimate the incidence of poliomyelitis and neonatal tetanus in Gujarat*. Gandhinagar, EPI Section, Directorate of Health and Medical Services, Government of Gujarat, 1981.
8. *Survey to estimate the incidence of poliomyelitis and neonatal tetanus in Tamil Nadu and Pondicherry*. Madras, Directorate of Public Health and Preventive Medicine, 1981.
9. *A preliminary report on the survey to estimate the incidence of poliomyelitis and neonatal tetanus in Andhra Pradesh*. Hyderabad, Department of Health and Family Welfare, Government of Andhra Pradesh, 1981.
10. Polio and neonatal tetanus survey in Maharashtra. *EPI bulletin (India)*, **5**: 38–53 (1982).
11. COCHRAN, G. G. *Sampling techniques*. New York, John Wiley and Sons, 1977.
12. HOAGLIN, D. C. ET AL. *Understood robust and exploratory data analysis*. New York, John Wiley & Sons, 1983.
13. BRINK, E. W. ET AL. Nutritional status of children in Nepal, 1975. *Bulletin of the World Health Organization*, **54**: 311–318 (1976).
14. BRINK, E. W. & NAKANO, J. H. Naturally acquired measles immunity in Nepal and Sri Lanka. *Tropical and geographic medicine*, **30**: 109–113 (1978).
15. BLACK, F. L. Measles antibody prevalence in diverse populations. *American journal of diseases in children*, **103**: 72–79 (1962).
16. HENDRICKSE, R. G. ET AL. Measles vaccination. Report of a large scale trial of further attentuated measles vaccine in Nigeria. *Journal of tropical medicine and hygiene*, **69**: 112–116 (1966).

Annex

HOUSEHOLD VISIT FORM

(1) Cluster number

(2) Date

(3) Team

(4) District

(5) Panchayat [village]

(6) Ward

(7) Location

(8) Household number ☐

(9) Head of household

(10) Total number of people in this household: ☐

List children by name. For death in previous 12 months or lameness, fill out special form.

(11) Born in previous 12 months (living)

Name	Birth date	Measles	Diphtheria	Pertussis	Tetanus
TOTAL ☐		☐	☐	☐	☐

(12) Born in previous 12 months (died)

Name	Birth date	Death date	Neonatal tetanus	Measles	Diphtheria	Pertussis	Tetanus
TOTAL ☐			☐	☐	☐	☐	☐

(13) Age 1–9 years (living)

Name	Age	Measles	Diphtheria	Pertussis	Tetanus	Lame (5–9 years only)
TOTAL ☐		☐	☐	☐	☐	☐

Total aged 5–9 years ☐

(14) Age 1–9 years (died in previous 12 months)

Name	Age	Measles	Diphtheria	Pertussis	Tetanus	Lame (5–9 years only)
TOTAL ☐		☐	☐	☐	☐	☐

Total aged 5–9 years ☐