International Journal of Epidemiology

Vol. 23, No. 6
Printed in Great Britain

# A Computer Simulation of Household Sampling Schemes for Health Surveys in Developing Countries

S BENNETT, A RADALOWICZ,* V VELLA** AND A TOMKINS†

Bennett S (Tropical Health Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel St., London WC1E 7HT, UK), Radalowicz A, Vella V and Tomkins A. A computer simulation of household sampling schemes for health surveys in developing countries. *International Journal of Epidemiology* 1994; **23**: 1282–1291.

*Background.* Cluster sample surveys of health and nutrition in rural areas of developing countries frequently utilize the EPI (Expanded Programme on Immunization) method of selecting households where complete enumeration and systematic or simple random sampling (SRS) is considered impractical. The first household is selected by choosing a random direction from the centre of the community, counting the houses along that route, and picking one at random. Subsequent households are chosen by visiting that house which is nearest to the preceding one.

*Methods.* Using a computer, and data from a survey of all children in 30 villages in Uganda, we simulated the selection of samples of size 7, 15 and 30 children from each village using SRS, the EPI method, and four different modifications of the EPI method.

*Results.* The choice of sampling scheme for households had very little effect on the precision or bias of estimates of prevalence of malnutrition, or of recent morbidity, with EPI performing as well as SRS. However, the EPI scheme was inefficient and showed bias for variables relating to child care and for socioeconomic variables. Two of the modified EPI schemes (taking every fifth house and taking separate EPI samples in each quarter of the community) performed in general much better than EPI and almost as well as SRS.

*Conclusions.* These results suggest that the unmodified EPI household sampling scheme may be adequate for rapid appraisal of morbidity prevalence or nutritional status of communities, but that it may not be appropriate for surveys which cover a wider range of topics such as health care, or seek to examine the association of health or nutrition with explanatory factors such as education and socioeconomic status. Other factors such as cost and the ability to monitor interviewers' performance should also be taken into account.

Cluster sample surveys are frequently used for the assessment of the health status of communities in developing countries. A sample of communities is selected, perhaps in several stages, and within each selected community a sample of households is selected.[1] Selection of households within a community should ideally be at random, and in practice this is most closely achieved by systematic selection from a numbered list of households. In many situations, however, there is no list or map of households available, and if the investigator does not have the resources to completely enumerate and map all the households in the community, some compromise method must be used.

A common alternative approach is the EPI household sampling scheme, developed by the World Health Organization's Expanded Programme on Immunization[2] for estimating vaccination coverage. In this procedure, in a rural community, the first household is selected by choosing a random direction from the centre of the community, counting the houses along that route, and picking one at random. If this household contains a child in the target age range (usually 12–23 months) it is included in the sample. The procedure for visiting subsequent households is to choose that house which is nearest to the preceding one. This process continues until the required number of individuals in the target range have been recruited.

This sampling procedure is simple to carry out, needing no mapping or listing of households, but has the disadvantages that the first house is chosen by a procedure that is somewhat biased, and that the sample is concentrated in one part of the community. These may not be problems if the individuals sought are found only relatively infrequently. A child in a narrow age range

* Tropical Health Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel St., London WC1E 7HT, UK.
** UNICEF, Uganda Country Office, PO Box 7047, Kampala, Uganda. Current address: SA3PH, The World Bank, 1818 H St. NW, Washington, DC 20433, USA.
† Centre for International Child Health, Institute of Child Health, 30 Guilford St., London WC1N 1EH, UK.

such as 12–23 months may be found only once in every four households, resulting in the sample being spread more widely, but a subject who is more common, say a child aged 0–4 years, or a woman aged 15–44, may be found in almost every household, resulting in a sample which is very tightly concentrated about the initial house.

A sample selected in this way may not be representative of the entire community. If such bias is consistent between communities, it will lead to bias in the overall estimates of prevalence, coverage etc. If not, it will lead to a decrease in the precision of these estimates, leading to wider confidence intervals, and to less power to detect significant differences between subgroups.

The validity of the EPI method has been evaluated for its original purpose of estimating immunization coverage,[3,4] but not in other contexts such as maternal and child health, nutritional status or other aspects of primary health care (PHC), where the age range of the individuals studied is likely to be wider. Amendments to the scheme aimed at making it more representative have been suggested[1,5] such as taking every fifth house, dividing each community into quarters and selecting a separate EPI sample from each one, or taking part of the sample from the centre of the community and part from the periphery.[6]

In this paper we use computer simulation to select repeated samples from a rural Ugandan population using simple random sampling (SRS), EPI sampling and four adaptations of the EPI scheme. We evaluate the effect of these sampling schemes on the bias and precision of estimates of a range of indicators of child health, nutritional status, health care and associated socioeconomic factors. We discuss the implications of our findings for the conduct of future surveys of this kind. A review and more general appraisal of the EPI method has appeared elsewhere.[7]

## METHODS

### *The Survey and the Data*

In March and April 1988 a baseline survey was carried out in Mbarara District in south-west Uganda in preparation for a UNICEF/Uganda Ministry of Health PHC project in that area. Thirty communities (villages) were selected and a complete census of each village taken. A household was defined as all those sharing a common cooking pot. These villages contained a total of 2532 households, ranging from 51 to 153 per village, and 4320 children under 5 years (range 86–238 per village) of whom data on 4129 (96%) were included in this study. For each household, information was collected on socioeconomic factors, water supply and sanitation etc. Each child under 5 years was measured for weight, length (<36 months) or height (>36 months), and mid-upper arm circumference (MUAC) according to standardized methods, and his or her morbidity for the past 2 weeks was recorded. The prevalence of malnutrition, morbidity and socioeconomic indicators are described elsewhere.[8] All houses in the village were accurately mapped.[9]

The variables used in the computer simulation are shown in Table 1. For the purpose of this study each variable has been treated simply as a dichotomy. Not more than 6% of the data were missing for any variable.

### *The Sampling Schemes*

In immunization coverage surveys[2] seven children are sampled from each community. To represent this design and the larger sample sizes used elsewhere, we took samples of size 7, 15 and 30 children from each community. The following schemes for the selection of households within each community were simulated:

(i) SRS: Simple random sampling. Each house in the community is given a unique number from 1 to n (where n is the total number of houses). A sample of houses is then selected using a table of random numbers.

(ii) EPI: The EPI method.
  (a) Selection of the first household: The investigator stands at a central point in the community and chooses a direction at random (e.g. by spinning a pen in the air and seeing how it lands). S/he counts the houses between the centre and the edge of the community in that direction. One of these houses is selected at random.
  (b) Selection of subsequent households: The investigator chooses the house whose door is nearest to the door of the household s/he has just left.

(iii) EPI3: The first household is selected by the EPI strategy (ii,a) above. Subsequent houses are selected by choosing a random direction and selecting the third nearest house in that direction.

(iv) EPI5: As (iii), but the fifth nearest household is selected.

(v) QTR: The community is divided into four quadrants; the EPI strategy (ii) is then used independently to select a quarter of the sample from each quadrant, starting at a central point in each quadrant.

(vi) PERI: Half of the sample is selected at the centre of the community and half at the periphery. A random direction is taken from the centre, and the *first* house in that direction is visited. Half of the sample is selected by visiting in turn the nearest households as in (ii,b). The investigator then returns to the centre, chooses again a random direction and visits the *last*

TABLE 1 *Variables considered in the simulation study*

| Variable | Meaning | Prevalence % |
|---|---|---|
| Nutrition | | |
| Height for age | Height for age z-score <–2.0 | 32.0 |
| Weight for age | Weight for age z-score <–2.0 | 18.1 |
| Weight for height | Weight for height z-score <–2.0 | 3.7 |
| MUAC | Mid-upper arm circumference <13.5cm | 20.7 |
| Morbidity | | |
| Fever | Child had fever in previous 2 weeks | 7.9 |
| Diarrhoea | Child had diarrhoea in previous 2 weeks | 8.8 |
| Respiratory infection | Child had respiratory infection in previous 2 weeks | 16.2 |
| Health care | | |
| Breasfed | Child is not currently breastfed | 66.5 |
| Pregnant | Mother is pregnant | 15.6 |
| Weighed in last 3 months | Child was not weighed in previous 3 months | 92.4 |
| Growth chart available | Growth chart is not available | 64.8 |
| Interpret chart | Mother cannot interpret growth chart | 87.8 |
| Scoioeconomic | | |
| Mother's education | Mother has ≥1 years of education | 50.4 |
| Father's education | Father has ≥1 years of education | 74.3 |
| Ethnic group | Father's ethnic group is not Banyankole | 19.9 |
| Religion | Father is Protestant | 54.9 |
| Subsistence farmer | Father is a subsistence farmer | 41.9 |
| Keeps cattle | Father keeps cattle | 19.6 |
| Grows a cash crop | Father grows a crop for sale | 12.2 |

house in that direction. The remainder of the sample is selected, again as in (ii,b).

If there was more than one child aged 5 in a household, then all were included in the sample. Households were sampled from the community until the required sample size (in terms of children) was achieved or surpassed. Each sampling scheme was simulated 1000 times for each sample size.

*Measures of Effectiveness*

For each simulation of the sampling procedure, we estimated the sample prevalence of each of the attributes listed in Table 1. We used the following measures[4] to summarize the performance of the various sampling schemes:

BIAS: The mean value of the sample prevalence over all 1000 simulations minus the expected prevalence under SRS. (In taking an unweighted SRS of equal size from each community, the expected prevalence is the mean of the 30 community population prevalences.)

VARIANCE: The sample variance of the 1000 sample means. This estimates the expected variance of the mean of a single sample, and its square root estimates the standard error of the mean.

MSE: Mean square error, equal to bias$^2$ + variance; a measure of the total error.

HDEFF: 'Household design effect'; the ratio of the variance for the given sampling scheme to the corresponding variance achieved under SRS of households. This will be one component of the 'design effect'[10] (*deff*) which measures the increase in variance of the complete sampling design, including stratification and cluster sampling, compared to an SRS of individuals taken from the entire region. Confidence intervals for the population mean will be wider by a factor $\sqrt{hdeff}$ than if households had been sampled by SRS, and sample sizes should be larger by a factor *hdeff* to compensate for this.

Values of bias, variance and MSE in the Tables are presented on the same percentage scale as the prevalences in Table 1.

*Implementation of the Sampling Schemes*

Maps of each village were digitized manually, and the x,y co-ordinates of each household added to its record. Random numbers were generated from the computer's internal clock. Sampling was without replacement, so that no household could appear twice in the same sample.

To select the first household in the EPI and associated schemes the centre of the village was defined by the median x and y co-ordinates of all the houses in the village. From this point a random direction was generated. All houses within a fixed short distance of a line drawn in this direction were considered to be on the 'path'. The number of houses on this path was counted, and one selected at random. Subsequent households were chosen by selecting that house which was nearest, and which had not been selected before in this sample.

## RESULTS

For reasons of space, results are shown only for sample sizes 7 and 30; those for sample size 15 are mentioned in the text.

### *Variance and Household Design Effect*

Table 2 shows that for sample size 7, there is little difference in variance between SRS, EPI3, EPI5 and QTR, with most of the *hdeffs* very close to one. The variances of the PERI scheme are much reduced compared to SRS, 10 of the 19 variables having *hdeffs* of less than 0.8. However for EPI sampling high *hdeffs* are shown for a few variables, notably ability to interpret a growth chart and father's ethnic group, the former showing an increase in variance of 36%. *Hdeffs* for nutritional variables are mostly less than one for the EPI scheme.

For sample size 30 (Table 3), the *hdeffs* for EPI3, EPI5 and QTR are again very close to unity (except for possession of a growth chart). EPI sampling appears to be more efficient than SRS for measuring nutritional status and morbidity, but its inefficiency is increased for other variables, with *hdeff* over 1.2 for six variables. The *hdeffs* for PERI are still low, but higher than for sample size 7. Untypically, EPI sampling provides particularly low variances for the prevalence of breastfeeding in all sample sizes.

The picture for sample size 15 (data not shown) is similar to that for sample size 30. Variances are even larger for EPI, with seven variables having *hdeff* over 1.2 (1.40 for interpretation of growth chart), and closer to unity for nutritional and morbidity variables. Variances are again small for PERI, and the other sampling schemes again have variances close to those of SRS, except for possession of a growth chart.

EPI5 generally has lower variances than EPI3 for the larger sample sizes, with QTR looking the best of the three when the sample size increases to 30.

For no sample size, nor for any sampling scheme, are large *hdeffs* seen for any of the nutritional status variables (height for age, weight for age, weight for height and mid-upper arm circumference) or for morbidity (fever, diarrhoea or respiratory infection in previous two weeks), mother's or father's education, breastfeeding or pregnancy status. It is socioeconomic and cultural variables (being a subsistence farmer, growing a cash crop, father's religion and ethnic group) and health care variables (whether child had been weighed recently, availability and interpretation of growth chart) that are affected most by the choice of sampling scheme.

### *Bias and Mean Square Error*

Table 4 shows bias and mean square error for the smallest sample size. The PERI scheme frequently leads to considerable bias, the magnitude being over 3% for possession of cattle, and frequently over 2% (note that this is an absolute, not relative, percentage bias). The bias of SRS is close to zero and for EPI3, EPI5 and QTR is almsot always less than 1%, but EPI leads to a bias of more than 1% for seven variables. The biases of all the sampling schemes become slightly smaller for sample size 15 (data not shown), but large biases are still seen for many variables with PERI, and for some variables with EPI. For sample size 30 (Table 5) biases are smaller again, although a few high values are seen for PERI and EPI.

Biases are most extreme for possession of cattle, and are also high for father's ethnic group, religion and level of education, and for mother's education and being a subsistence farmer for the smaller sample sizes. Interestingly, PERI, EPI, EPI3 and to a lesser extent EPI5, overestimate the prevalence of low height for age whatever the sample size.

The effect of bias and variance is combined in the notion of mean square error (MSE). For sample size 7 (Table 4) EPI sampling shows a high MSE for interpretation of a growth chart, father's ethnic group, religion and level of education, being a subsistence farmer and keeping cattle. The combination of low variance and high bias for PERI sometimes results in a low MSE (weight for age, height for age, breastfeeding and availability of a growth chart), and sometimes in a high MSE (fever, ethnic group, pregnancy and keeping cattle). The MSE for EPI3, EPI5 and QTR are broadly similar to those for SRS.

For sample size 15 (data not shown), EPI sampling shows MSE more than 20% above that of SRS for nine of the 19 variables, and PERI for eight. EPI3 (with 5), EPI5 (3) and QTR (2) also show some high values. Low values are almost non-existent. The picture for sample size 30 is very similar (Table 5), but note the very large MSE for religion and keeping cattle with EPI compared to those with SRS.

TABLE 2 *Variance of different sampling schemes with sample size 7, and (below) household design effect (hdeff), or ratio to srs variance*

| | Sampling scheme [a] | | | | | |
|---|---|---|---|---|---|---|
| | SRS | EPI | EPI3 | EPI5 | QTR | PERI |
| Nutrition | | | | | | |
| Height for age | 11.9 | 10.5 | 11.1 | 11.1 | 10.7 | 8.6 |
| | | 0.88 | 0.93 | 0.93 | 0.90 | 0.72[c] |
| Weight for age | 7.5 | 6.6 | 7.7 | 7.4 | 7.2 | 5.2 |
| | | 0.88 | 1.03 | 0.98 | 0.96 | 0.69[c] |
| Weight for height | 1.8 | 1.5 | 1.8 | 1.8 | 1.4 | 1.1 |
| | | 0.84 | 1.00 | 1.00 | 0.81 | 0.60[c] |
| Mid-upper arm circumference | 6.7 | 7.1 | 7.8 | 6.8 | 7.0 | 6.0 |
| | | 1.06 | 1.17 | 1.02 | 1.05 | 0.91 |
| Morbidity | | | | | | |
| Fever | 3.5 | 3.6 | 3.8 | 3.8 | 3.2 | 2.9 |
| | | 1.02 | 1.09 | 1.09 | 0.92 | 0.81 |
| Diarrhoea | 3.8 | 3.5 | 3.4 | 3.5 | 3.4 | 3.1 |
| | | 0.94 | 0.91 | 0.94 | 0.91 | 0.82 |
| Respiratory infection | 7.6 | 8.1 | 7.7 | 6.8 | 6.9 | 6.9 |
| | | 1.06 | 1.02 | 0.90 | 0.92 | 0.91 |
| Health care | | | | | | |
| Breastfed | 7.8 | 5.4 | 7.0 | 7.2 | 7.1 | 4.7 |
| | | 0.69[c] | 0.91 | 0.93 | 0.92 | 0.61[c] |
| Pregnant | 9.1 | 9.7 | 9.6 | 8.7 | 8.2 | 7.4 |
| | | 1.07 | 1.06 | 0.96 | 0.90 | 0.81 |
| Weighed in last 3 months | 4.8 | 5.5 | 4.9 | 4.5 | 4.1 | 3.6 |
| | | 1.14 | 1.02 | 0.94 | 0.85 | 0.75[c] |
| Growth chart available | 13.5 | 14.7 | 13.6 | 13.9 | 15.0 | 10.9 |
| | | 1.09 | 1.01 | 1.02 | 1.11 | 0.80 |
| Interpret chart | 7.8 | 10.6 | 8.1 | 7.4 | 7.3 | 6.4 |
| | | 1.36[b] | 1.04 | 0.94 | 0.94 | 0.82 |
| Socioeconomic | | | | | | |
| Mother's education | 17.3 | 17.8 | 18.8 | 18.4 | 17.9 | 11.1 |
| | | 1.03 | 1.09 | 1.06 | 1.03 | 0.64[c] |
| Father's education | 15.1 | 15.5 | 14.5 | 15.4 | 13.9 | 11.7 |
| | | 1.03 | 0.96 | 1.02 | 0.92 | 0.78[c] |
| Ethnic group | 9.1 | 11.3 | 9.0 | 8.4 | 9.6 | 6.9 |
| | | 1.25[b] | 0.99 | 0.92 | 1.06 | 0.76[c] |
| Religion | 15.8 | 18.4 | 17.3 | 15.7 | 17.5 | 12.7 |
| | | 1.16 | 1.10 | 0.99 | 1.11 | 0.80 |
| Subsistence farmer | 18.7 | 21.1 | 19.5 | 20.0 | 18.7 | 14.9 |
| | | 1.13 | 1.04 | 1.07 | 1.00 | 0.80[c] |
| Keeps cattle | 11.9 | 11.3 | 11.7 | 12.9 | 11.4 | 8.1 |
| | | 0.95 | 0.98 | 1.09 | 0.97 | 0.68[c] |
| Grows a cash crop | 8.3 | 8.3 | 7.3 | 8.5 | 7.6 | 6.8 |
| | | 1.00 | 0.88 | 1.03 | 0.92 | 0.82 |

[a] For an explanation of the sampling schemes see text.
[b] *hdeff* >1.2
[c] *hdeff* <0.8