Evaluation of a Standardized Survey Design Proposed for Use in Epidemiological Research on AIDS

DONALD R HARRIS,* STANLEY LEMESHOW,* STEVEN K LWANGA,** JAMES CHIN†, AND ROGER DUCHESNEAU‡


A Monte Carlo model simulating actual populations was employed to evaluate the precision in estimation of a standardized sampling method proposed by the Global Programme on AIDS of the World Health Organization, for general use in collecting population-based data on HIV seroprevalence. It appears that in real populations, where there is likely to be "pocketing" of infection, the proposed methodology will generally fail to provide estimates accurate to within 1% of the true population value. However, if the primary objective of a particular survey is to construct confidence intervals that include the true population HIV seroprevalence rate, then this survey method appears to be a reasonable choice. This study also suggests that selection of only one adult per household improves the precision of resulting estimates. However, since selection of only one adult per household would require that more households be visited, any gain in precision would need to be weighed against the likely increase in cost of household visits.

Public health programmes require reliable information on the distribution and prevalence of infection to the human immunodeficiency virus (HIV)—the aetiological agent for acquired immunodeficiency syndrome (AIDS)—in order to develop and direct prevention/control activities against this agent. Serological surveys designed to obtain this information must use sampling methods that are affordable and reliable. Probability proportionate to size (PPS) cluster sampling is one such widely applied sampling technique.

In general, PPS sampling begins by subdividing a population into clusters for which reasonably accurate estimates of the total population size are known. At the first stage of sampling, a subset of clusters is selected in such a manner that the probability of a particular cluster being selected is directly proportional to the total number of individuals in that cluster. With this approach, densely populated clusters have a higher probability of being selected than sparsely populated clusters and may be selected more than once for inclusion. At the second stage of sampling, a random sample of enumeration units is selected from each of the clusters chosen.

Certain features of classic PPS cluster sampling make it difficult to implement in the field. In particular, it may be impossible to randomly select households at the second stage of sampling, especially in developing countries. Therefore, since 1978, the Expanded Programme on Immunization (EPI) of the World Health Organization (WHO) has advocated a modification of PPS cluster sampling for surveys designed to estimate levels of immunization coverage among children that involves a second stage without random selection.

A computer simulation study which evaluated the EPI methodology versus classical PPS cluster sampling for use in estimating vaccination coverage concluded that the technique yielded reasonably accurate overall population estimates for typical populations. However, this study warned that the EPI sampling scheme would not always satisfy stated objectives with respect to precision in estimation. This was demonstrated with a 'worst-case' population constructed with high population density, low levels of vaccination coverage and pocketing of vaccination in all clusters.

More recently, the WHO Global Programme on AIDS (GPA) developed guidelines for estimating the
prevalence of HIV infection in large populations in areas where the predominant mode of HIV transmission is heterosexual intercourse and where the prevalence of HIV infection in large cities is at least 1% or more among sexually active adults. The sampling strategy proposed by WHO represents a modification of the EPI survey method aimed at addressing concerns raised by the study of Lemeshow et al. 3

The original EPI method was not developed to estimate small proportions or provide estimates with a high degree of accuracy. In an effort to increase precision in estimation, the EPI method was modified for use in HIV serosurveys to include the selection of a larger number of clusters at the first stage of sampling and a larger number of adults during subsequent sampling.

Given the uncertainty as to whether these modifications would result in estimates with the desired level of precision, the current investigation was designed to evaluate the proposed method with respect to bias and variability in estimation of the true population HIV seroprevalence rate relative to classical theoretical PPS cluster sampling employing simple random sampling (SRS) of households at the second stage of sampling. The investigation was based on implementation of a Monte Carlo model designed to simulate actual populations and each stage of survey sampling in order to assess precision in estimation of the proposed methodology.

METHODS

Creation of Population Clusters

In the simulation design up to 100 population clusters could be generated, each representing a community of approximately 4800 individuals, half of whom are over 15 years of age (each cluster encompassed a total of approximately 600 households, each with an average of four adult occupants). For the purpose of this investigation, individuals aged 15 and older were considered to be sexually active adults and therefore at risk of having HIV infection.

Four different population densities were defined: (a) low density clusters representing rural areas with scattered population; (b) moderate density clusters representing more highly populated areas; (c) high density clusters representing highly populated areas; and (d) very high density clusters representing highly urbanized areas.

In actual populations, households are usually not distributed in a random fashion. Instead, households may be clustered around a central gathering point such as a market place or church, or may be more likely to be built along a road, stream or river. To reflect these differences, four basic household distribution patterns were included with probabilities of household placement set accordingly.

One of the household distribution patterns was designed to randomly place households throughout the cluster. Another option, designed to simulate a central road or stream, concentrated household placement linearly through the middle of the cluster. A third possible household pattern was designed to represent a community with a single, central focal point for congregation of dwellings. The last household distribution pattern assumed that there were two population foci in the community.

The total number of adult occupants was randomly assigned to each household in such a way that the average number per household was four; the specific algorithm used in assigning the total number of adults to each household employed the random number generator developed by Wichmann and Hill. 4

Specific rates of HIV infection, generally ranging between 1 and 15% or more, were designated for each type of cluster depending upon the nature of the area represented. The infection status of individual adults was identified according to the specific rate and pattern of infection. Actually, two different strategies were used to determine which adults were seropositive for HIV infection. Using the first strategy, infected adults were identified at random among all adults in a particular cluster in accordance with the designated rate of infection. The second strategy established one or more ‘pockets of infection’ with specified rates of infection.

Clustering of infected adults within households was simulated as follows. For households having more than one adult occupant, the first two adults were considered sexual contacts. If one member of this pair was found to be infected, the determination of the infection status of the other was based on a higher probability of infection.

Having designated the overall infection rate and the higher rate applied to sexual contacts, the baseline infection rate was determined by the computer using an algorithm based on a binomial distribution for infection.

Actual household surveys are almost certain to include individuals who are either unavailable or unwilling to participate. This can be particularly important since high levels of non-response by particular subgroups can bias the results of a survey to a considerable degree. For each adult, a variable indicating their participation status was included in the simulation in order to evaluate the influence of non-response upon estimates obtained using the two sampling
strategies. Although the assignment of response status to individuals was based on a random process, different probabilities of non-response were used for infected and non-infected adults.

**Sampling Process**

To enable the simulation of field applications of these sampling methods, five different populations of 100 clusters were constructed with parameters chosen to reflect actual population characteristics of African nations. The first population, which will hereafter be referred to as the 'Standard' population, was designed to represent a predominantly rural country. A second predominantly rural population ('Rural') was created with differing cluster characteristics. The 'Mixed' population, was designed to simulate a fairly even blend of rural and urban areas. The two remaining artificial populations, created to simulate predominantly urbanized areas, are referred to as the 'Urban' and 'City' populations. The distribution of clusters according to population density is presented in Table 1 for each of the five study populations.

<table>
<thead>
<tr>
<th>Population</th>
<th>Very high</th>
<th>High</th>
<th>Moderate</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>5</td>
<td>10</td>
<td>30</td>
<td>55</td>
</tr>
<tr>
<td>Rural</td>
<td>5</td>
<td>10</td>
<td>70</td>
<td>15</td>
</tr>
<tr>
<td>Mixed</td>
<td>16</td>
<td>37</td>
<td>30</td>
<td>17</td>
</tr>
<tr>
<td>City</td>
<td>60</td>
<td>30</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Urban</td>
<td>26</td>
<td>37</td>
<td>27</td>
<td>10</td>
</tr>
</tbody>
</table>

According to the EPI-like strategy, the second stage of sampling begins within a chosen cluster with the selection of a single household at random to serve as the starting point for the survey. Having visited this household and collected the desired information on all adult residents, the physically-closest household is then identified for inclusion in the survey using a specific algorithm. This process of visiting physically-closest households continues until the desired total number of adults within the cluster have been surveyed.

In contrast, the SRS strategy randomly selects households (without replacement) in each cluster until the infection status of the required number of adults has been recorded. For example, collection of information on a total of 200 adults would require that 50 households be selected at random (assuming an average of four adults per household). However, if only one adult per household were to be surveyed then 200 households would need to be randomly chosen.

**Statistical Analyses**

Assessment of bias and variability in estimates of the population seroprevalence rate were of primary importance in comparing the two survey methods. In order to minimize the impact of sampling error in this evaluation, this assessment was based on 500 replications of the sampling from each population. Each replication began with the selection of an original PPS sample of clusters at the first stage of sampling followed by selection of 200 adults using each of the comparison methods.

Each replication of the sampling provided an estimate of the proportion of adults with HIV infection in the entire population. Specifically, the expected value of the population HIV seroprevalence rate was estimated as the average value of the proportion of adults infected with the AIDS virus over all of the clusters sampled in a single replication of the sampling process.

A variety of measures of precision in estimation were computed for the EPI-like methodology which were then compared to those computed based on simple random sampling at the second stage of sampling. These included the bias, absolute relative bias (ARB), and the mean square error (MSE) for the estimated population HIV seroprevalence rate. In addition, the ability of the EPI-like sampling strategy to meet certain stated objectives with regard to precision in estimation was assessed. These objectives included whether an estimate was within a certain number of percentage points of the true value of the parameter, whether estimates fell within the more restrictive interval defined by the actual value plus or minus some percentage of its value, and lastly, whether the confidence interval constructed about the estimate included the true population parameter.

Bias as a statistical concept represents the difference between the estimate and the actual value of the parameter. The magnitude of the bias provides some indication of the precision of the estimate. However, since the bias is not invariate to the magnitude of the parameter being estimated, the ARB is often preferred as a measure of precision in estimation. The ARB expresses the bias as a proportion of the true value of the parameter, thereby providing a direct measure of the precision of an estimate which is independent of the magnitude of the parameter estimated. Accordingly, large values of the ARB tend to be associated with imprecise estimates while small values are indicative of greater precision in estimation. Lastly, the MSE was also used to assess the precision of the EPI-like strategy relative to SRS in estimating HIV seroprevalence. Computational procedures and
RESULTS

Results of a single execution of the simulation program developed for this study may include a variety of summary tables. Certain tables are included which assess the precision of each method of estimation with respect to the true overall parameter based on the entire population of 100 clusters. In such cases, poor performance could be indicative of a lack of precision in estimation, or could possibly reflect selection of an unrepresentative PPS sample of clusters on which to base estimates for the population as a whole, or may result from a combination of these two factors. To avoid potential for bias in the selection of clusters, results were based on selection of 500 different PPS cluster samples.

Appendix 1 presents typical results of estimation of the true population seroprevalence rate for selected repetitions of the sampling from a single population; the actual value of the population seroprevalence rate was 8.2%. The population estimates based on the EPI-like and SRS strategies were computed for each replication as a simple average of the individual rates estimated for the 50 clusters selected. The ARB, the mean square error and the bias are displayed in the remaining columns of the table.

In general, the values for the measures of precision were smaller for SRS. While the ARB associated with SRS estimates seldom exceeded 20%, values of 30% or more were observed with the EPI-like sampling scheme. The reported values of the MSE tended to be small for both methods of sampling, especially for the SRS strategy.

An estimate of the true population infection rate was also derived for both methods of sampling as an average of the individual replication estimates computed over the total number of repetitions of the sampling. This value is reported in the last row of the Table, labelled, ‘Overall’, along with the ARB, MSE and the bias. The findings based on this overall estimate of the true population parameter were similar to those described for the body of the Table.

Rather than attempt to condense the vast quantity of information contained in similar tables summarizing results of estimation of the population parameters for each of the five study populations, the performance of each sampling strategy was judged on the basis of its ability to meet specific stated objectives. Results based on sampling from these populations constructed with and without pocketing of infection and employing selection of all or only one adult per household are included. Separate tables are utilized in presenting results based on responders only. The results of comparison of the two sampling methodologies are summarized for all study populations in Tables 2-7.

Table 2 displays the number of sample estimates accurate to within one percentage point of the actual HIV infection rate. The results provided for each population constructed without pocketing of infection indicate the total number of estimates within the specified range of the true population parameter to be very similar for the two methods of estimation. However, with inclusion of pocketing, the number of SRS estimates within the range of interest greatly exceeded that of the EPI-like method. The number of estimates meeting this objective in precision in estimation tended to increase when based on selection of one adult at random from each household visited.

The total number of sample estimates falling within the more restricted range defined by the true infection rate ±5% of the!true value is shown in Table 3. There was little difference between the survey methods in the absence of pocketing even under this more restrictive condition. However, when pocketing of infection was included, the classical SRS strategy performed substantially better than the EPI-like method. In general, precision in estimation improved with both sampling methods when based on random selection of only one adult per household.

Table 4 presents the total number of 95% confidence interval estimates that included the true population parameter. The observed data fail to demonstrate a significant difference between the two methods regardless of population type or whether pocketing of infection was included. In this case, the effect of selection of one versus all household adults was equivocal.

Tables 5 to 7 present results of estimates from 500 PPS samples of clusters in which non-responders were excluded from the analyses.

The number of sample estimates accurate to within one percentage point of the actual HIV infection rate are shown in Table 5. In the absence of pocketing of infection, the EPI-like method performed best when selection was based on all household adults. However, if only one adult per household was selected then SRS was better able to meet this stated objective. With the inclusion of pocketing of infection, SRS tended to perform better when all household adults were selected and the EPI-like method when only one adult was randomly chosen from each household. The differences between the two methods were, however, quite small.