EPIDEMIOLOGY 1, 2, 3: STUDY AND SAMPLE DESIGN

MOHAMED H. HUSSEIN, MEDHAT K. EL-SAYED, MAHA TALAAT, AMAL EL-BADAWY, AND F. DEWOLFE MILLER
Faculty of Medicine, Cairo University, Cairo, Egypt; Faculty of Medicine, Suez Canal University, Ismailia, Egypt;
Department of Community Medicine, Theodor Bilharz Research Institute, Cairo, Egypt; Faculty of Medicine, Zagazig University,
Zagazig, Egypt; School of Public Health, University of Hawaii, Honolulu, Hawaii

Abstract. Bad sample designs and selection bias have plagued studies on schistosomiasis, and as a result some believe that schistosomiasis is too focal, making it difficult to draw reliable samples. The Epidemiology 1, 2, 3 (EPI 1, 2, 3) sample design, although complex, demonstrates that sampling theory is readily applicable to epidemiologic studies of schistosomiasis. The EPI 1, 2, 3 sampling scheme was designed to achieve the smallest feasible standard errors given EPI 1, 2, 3 objectives and certain logistical constraints. The sample design is a multi-stage selection of villages (ezbas, which were stratified by size) and households within each of 9 purposely selected Egyptian governorates. Villages and households were systemically selected from census frames. The sampling of ezbas was especially difficult because of the lack of complete sampling frames and their wide variation in population size. Ultimately, ezbas were stratified by size and then randomly selected from each stratum. Sample sizes for villages and ezbas and individuals within ezbas were calculated based on EPI 1 and 2 objectives, respectively. No re-selection was made for non-respondents. A 20% subsample of the full sample was drawn for clinical and ultrasonographic examinations.

The sample selected from individual governorates closely parallels the age structure of the 1986 census of the respective rural populations. Details of the study design and related methods are given below.

Epidemiology 1, 2, 3 (EPI 1, 2, 3) was an epidemiologic investigation of Schistosoma haematobium and/or S. mansoni infections in humans. The study methodology was designed and developed specifically to address the 3 EPI 1, 2, 3 objectives.1

Rather than examine an entire rural population, a sample of the defined target population was needed for the study. In turn, the sample estimates were used to make inferences about the whole target population. With notable exceptions, sample designs for studies on schistosomiasis in Egypt typically have been limited to single endemic villages. Only 1 study included more than 1 village and reported unbiased sampling estimates of schistosome infection for a defined target population.2 Typical results from other studies did not adjust or account for differential probabilities of selecting villages in calculating overall prevalence estimates3 or provide sufficient detail on sampling methods or both.4,5 Others have commented that schistosomiasis is too focal, preventing the use of standard sampling methods.6 The sample design by Farooq and others2 avoided these problems by using an equal probability of selection to each element (EPSEM) sample design that gave equal probability of selection to each element in the sample. Most useful to this study was the sampling methodology and the details of sampling techniques for application in rural Egypt used by Farooq and others2 in the Egypt 49 studies located in the Nile Delta. The report on the evaluation of the Upper Egypt National Schistosomiasis Control Program was also useful for establishing sampling frames for villages and defining sampling strata for rural Egypt at a governorate level.7

A good sample design with an adequate sample size will produce accurate estimates and precise standard errors (SEs). In absence of SEs, errors, the precision of prevalence (or incidence) estimates is unknown. Not knowing the precision of estimates undermines the analysis and determination if the study objectives have been achieved. However, no surveys of schistosomiasis in Egypt were found in which SEs were reported. The SEs for studies designed by Farooq and others2 would have been relatively easy to determine, but calculating SEs from more complex samples with varying population sizes, clusters of unequal sizes, can be complex and time-consuming.7

Our purpose here is to detail the study and sample design methods used throughout EPI 1, 2, 3 and provide an evaluation and discussion of these methods. The 3 EPI study objectives were originally conceived as independent nationwide studies. As mentioned in the introduction of this supplement,1 there were strong logistical and administrative justifications for combining the 3 objectives into 1 overall multi-center project. An overall sample design was developed to maximize the integration of the 3 different objectives with an emphasis on EPI 1. For EPI 3, the time-consuming ultrasonographic examination necessitated using a smaller sample size.

In regard to study measurements, more recent studies in Egypt and elsewhere provide a basis for choosing appropriate methods for the detection of schistosome ova in the urine and stool8,9 and the application of ultrasonography to population-based studies.10,11 Details on the application of these methods in EPI 1, 2, 3 have been provided separately.11,12

MATERIALS AND METHODS

Operationalization of the objectives. Fundamental in developing the design was the need for unbiased representative estimates of the prevalence of both S. haematobium and S. mansoni for an entire rural population in each of the selected governorates. The design could not exclude rural communities due to the presence or absence of health care facilities, control program activities, or because of remote or difficult accessibility. In fact, the study design specified the inclusion of very small villages (ezbas or tawaubes [satellites]) that have been frequently excluded or undefined in past surveys.1–4 The boundaries of the target population were defined as the rural population (further defined below) in each of the 9 separately selected governorates. Other requirements of the sample design dictated that both villages or ezbas and individuals be considered as units of observation in the analysis.

A trade off between sample size and funding, time, and
human resources had to be considered. The original plan for a nation-wide study of Egypt was reduced to 9 selected governorates from 17 rural governorates in the Nile Valley. The governorate selection was made by experts to sample governorates that would be representative for Egypt. All 9 governorates are known to be endemic for either or both species: 5 governorates were selected from the Nile Delta and the Suez Canal region and 4 more governorates were selected from Middle and Upper Egypt. Table 1 shows the selected governorates, the number of villages, the rural population as of the 1986 National Census as reported by the Central Agency for Public Mobilization And Statistics (CAPMAS; Nasr City, Egypt), and the number of subjects from that governorate included in EPI 1, 2, 3. Note that a village may include one or more ezbas. Figure 1 is a map of Egypt showing the location of the selected governorates.

Study design. The EPI 1, 2, 3 epidemiologic study design was a hybrid cross-sectional prospective study design of a probability sample from a defined rural population. Cross-sectional data from the sample provided estimates of point or period prevalence of infection, intensity of infection, and other related measures of schistosomiasis morbidity. The cross-sectional data describe the current distribution or patterns of infection and disease at the point in time (baseline) that the data were collected.

Sample design. Early on a number of sample designs considered EPSEM were excluded due to the need to have control over sample size at the village level, rather than be proportional to the village population size. Moreover, the multicenter nature of the project organization required decisions regarding the advantages and application of a uniform sample design to all select governorates. Some of the advantages of using a uniform sample design throughout was to reduce variance in SE estimates that might arise from different designs, standardize the rules and operations of the sample selection process, and standardize the analysis and computing of estimates and SEs. A single sample design was implemented to accommodate logistic, geographic, and administrative differences among the governorates.

A multi-stage sampling approach was chosen to minimize field cost and reduce time needed to travel between rural locations. The first stage primary sampling unit was the village. Detailed definitions were developed for village, ezba, house, dwelling, family, and individual. Before starting work in any village, village leaders and committees were contacted and details of the project explained. All village activities were coordinated with these village leaders and committees.

The sample frame for the first stage was a list of all rural villages and their respective populations for each selected governorate. This was obtained from available CAPMAS publications. The CAPMAS defines a village as a village administrative unit, which includes the central village and its surrounding ezba or tawaabe (satellites). The CAPMAS frame groups villages by district. Districts were randomly reordered and concatenated top to bottom to make a full list of eligible villages to select. Based on extensive prior field experience9,10 and the results of pilot studies in 1990 and 1991, villages with populations greater than 15,000 were redefined as urbanized (villages in which agriculture was no longer the predominant economic activity) and excluded. Villages were randomly selected.

Each selected village was mapped and each house was numbered with paint starting from 1. All ezbas associated with each village had to be enumerated because ezbas were not listed separately in the CAPMAS frame. Some villages had only 1 ezba, whereas others had more than 30. Ezbas were selected by a second-stage stratification method according to size (number of households). Each village and ezba was separated into 4 strata according to size for the second stage of the sample design with stratum 1 as the administrative central village. The second stage sample was drawn from each of 4 strata: stratum 1 = 500 or more households (village); stratum 2 = 300–499 households; stratum 3 = 50–299 households; stratum 4 = less than 50 households.

The third stage sample of households was stratum dependent. Households in strata 1, 2, and 3 were then randomly sampled in villages. All households of the selected clusters in stratum 4 were sampled.

The fourth stage sample included all individuals living in the selected household. No age limit was set and all members of the household were included even if traveling or working abroad at the time of the study. In some cases, the latter individuals returned at the time when specimens were being collected and were included, although the data from interviews were missing. Otherwise, these persons were considered as non-respondents.

All selected households and inhabitants were given a unique identification number and invited to participate in the study. Participation was entirely voluntary and followed the Egyptian Ministry of Health guidelines for informed consent. Examination of the participants began in May 1991.

Interview data. All participants were interviewed by trained field workers using pretested data forms that were translated into colloquial Arabic.

Parasitology. All villagers were asked to provide a urine and a stool specimen.12 All persons found positive for S. haematobium or S. mansoni were given free treatment with praziquantel (40 mg/kg). All egg counts were transformed to log10. The anti-log of the mean log10 of positive slides was used as the geometric mean egg count per gram of stool (GMEC).

Sample for physical examination and ultrasound examination. Cost, personnel, and time constraints for ultrasound examination have been considered. Only a 1 in 5 sample was examined ultrasonographically and by physical ex-
Figure 1. Map of Egypt showing the 9 governorates included in Epidemiology 1, 2, 3.
SCHISTOSOMIASIS IN EGYPT

amination. Ultrasound examination was performed on one-fifth (20%) of randomly selected households.

**Sample size.** To determine the number of villages and ezbas needed in the first stage and stage 4 (the number of individuals in households), the objectives for EPI 2 were used. The EPI 2 required a sufficient number of villages (including ezbas) to investigate village related factors and a sufficient number of individuals to have precise estimates of prevalence and intensity of infection at the village level. Sample size for villages was calculated using a regression model for detecting differences in continuous variables (village or ezba prevalence) and different sample sizes were estimated based on various assumptions of prevalence and variation between villages. The number of individuals necessary to sample in each village and ezba was determined using a prevalence estimate of 50%.

**Probability of selection.** The sample design provides a non-zero probability of being selected at each stage. The probability of the i<sup>th</sup> village being selected is expressed as P(V<sub>i</sub>) which is equal to 1/K, where K is the number of villages in the CAPMAS village frame for a given governorate. The probability of selection at the second stage strata of ezbas was P(E<sub>ij</sub>|V<sub>i</sub>), which was the conditional probability that E<sub>ij</sub> (for ezba) in the j<sup>th</sup> stratum was selected given that V<sub>i</sub> was selected. Finally, the probability of a household being selected in this stage was P(H<sub>ijk</sub> | E<sub>ij</sub>, V<sub>i</sub>) or the conditional probability that H<sub>ijk</sub> (for house) was selected given that E<sub>ij</sub> and V<sub>i</sub> were selected. Therefore, the overall probability for being selected in the sample from the entire rural population of a governorate was P(H<sub>i</sub> | E<sub>k</sub>, V<sub>i</sub>) or the probability that H<sub>i</sub> was selected. Since all persons in the selected house were selected, P(H<sub>i</sub> | E<sub>k</sub>, V<sub>i</sub>) is also the probability that each person living in the house is selected into the sample.

**Non-response.** Non-response is the percentage of persons choosing not to participate in the study. It was anticipated that some individuals may not be willing or able to participate in the study or may be away from the village at the time of the field operations. In some cases, the head of household would decline and subsequently prohibit any family member from participating. It was expected that the largest group of non-responders would be young children.

No replacement was made for any selected individual or household who did not participate. Data from the pilot studies (see below) indicated that non-response would be in the range of 10–15% of the sample. Accordingly, sample size selection for the number of households in any given village was increased by 25%.

Measures to reduce or minimize non-response were given to all field teams. When non-response was unavoidable, almost always the age and sex of the non-respondents was available from the village roster (census list) or could be obtained. Also the reasons for the non-response could frequently be obtained. It was stressed that anyone not participating because they were ill, bed ridden, or otherwise in a position that prevented their participation should be identified to the field team physicians for home visits.

The sample design dictated use of all data collected. That is if data on urine specimens are obtained on 1,000 persons but only 900 persons gave stool specimens, then the estimates for *S. haematobium* would be based on 1,000 and the estimates for *S. mansoni* would be based on 900.

**Data forms and questionnaires.** The design, testing, and reproduction of all data forms for the entire study was completed by the Core Team to ensure uniform data collection and recording. After a data form was finalized, it was printed in sufficient quantities and distributed to each of the 9 governorates teams. The study included the following forms:

- **Village roster form (Arabic).** This was a bound book in which the members of each household in the village were recorded by name, age, sex, family relationship, and EPI 1, 2, 3 project identification number. A separate book was used for each village and ezba.

- **Village form (Arabic).** One form was used to record the features of each village and ezba selected. Information included community water supply, sanitation, irrigation, and crops.

- **Household form (Arabic).** One form was completed for each selected and participating household. Data included source of household water and sanitation, size of house, and electric appliances.

- **Person form (Arabic).** One person form was completed for each selected participant. Data included age, sex, literacy, history of infection and treatment, and water contact.

- **Stool forms.** Three different forms were completed for each stool specimen received.

- **Urine forms.** Two forms were completed for each urine specimen received.

- **Clinical forms (English).** One clinical form for recording the results of a physical examination and other data was completed for each clinical examination.

- **Ultrasound form (English).** One form was used for recording the results of the ultrasound examination.

- **Pilot study.** All methods were fully tested and unforeseen problems were identified beforehand by conducting a full-scale pilot study in 1990. One village was selected by each of the investigators from their respective governorate. A systemic sample of village houses was drawn of sufficient number to include 500 or more household members. Data for all EPI 1, 2, 3 measures were taken, analyzed, and presented at an EPI 1, 2, 3 pilot study workshop held in Alexandria. The pilot study experience was invaluable in testing team members on the conduct of the study under village conditions, identifying logistical problems, improving participation rates, revising data forms, questionnaires, and census and sampling methods.

Following the pilot study phase, considerable effort was made in redesigning and retesting the data forms. A new field team was established solely to test revised forms. Repeated visits to nearby villages were made to test and retest revised data forms. This team in turn was used to train the field teams for the respective 9 governorates.

The EPI 1, 2, 3 manual for conducting field work for the full study. Following the pilot study, numerous revisions were made for recording data and interviewing. The final sample design and rules for sample selection of households and individuals for the field teams were completed. Quality control procedures for all measurements were finalized. Data entry procedures were standardized by software programming that initialized and setup each specific field team computer and configured the computers to be dedicated and restricted for EPI 1, 2, 3 data entry. These and all details for the implementation of the project in the field were organized.
and published in an *EPI 1, 2, 3 Manual for Conducting Field Work for the Full Study*. The manual had sections in both English and Arabic and was invaluable in evaluating on-going data collection in the field.

**Data analysis.** All data entry was verified using Epi-Info 5 with an Arabic format. Software written by the Core Team was used to configure the computers used by each field team for data entry. The configuration restricted the use of these computers to only data entry and verification. Since the sample selection was not proportional to the respective populations of each village or ezba, estimates of prevalence or GMEC could not be made without adjustments. Therefore, the probability of selection for each participant was calculated and converted to a population weight, i.e., 1 over the probability of being selected. Selection probabilities and the respective weights were also calculated to include estimates for non-response given that information was available for most non-respondents for age and sex and assuming similar age-sex specific infection patterns. Survey Data Analysis (SUDAAN; Research Triangle Institute, Research Triangle Park, NC) was used for calculating estimates of prevalence, GMEC and their respective SEs. This software incorporates the probability of selection and the sample design in estimating SE. We used the SUDAAN geometric means function to transform egg counts, estimate geometric means, and the geometric mean SE. Geometric means are not normally distributed, which complicates computing confidence intervals. The approximation used for confidence interval estimation was $\exp(SE/GMEC-Z_{1-\alpha} \pm \ln(GMEC))$ and it was assumed that sample means approach normality when sampled from a large population. Odds ratio was used to show an association between dichotomous variables and trend analysis (least squares method) for measuring the association between prevalence and GMEC at the village level. Chi-square was used to test statistical differences between villages. Otherwise, 95% confidence intervals were used to test for significant differences ($SE \pm Z_{1-\alpha} \cdot \text{estimate}$).

**RESULTS AND DISCUSSION**

A uniform sampling method was implemented for selection of the EPI 1, 2, 3 sample in 9 separate governorates. The sample size for each governorate and respective rural population are shown in Table 1. More details on the number of villages, households, and persons selected are provided in the subsequent reports on each of the 9 governorates. The collection and recording of all data for the study were done using the same methods throughout.

Figure 2 shows the age structure, by 5-year age groups, for the rural population of the Upper Egyptian governorate of Assiut (CAPMAS, 1986) compared with the age structure of the sample that was selected for Assiut. The rural population in the 0–4-year-old children was greater than the sample, indicating that the sample may under-represent this group. More recent results from demographic projections in Egypt have shown decreasing fertility in the rural areas over the past 10 years. Given decreasing fertility, the proportion of the 0–4-year-old children in our sample may be a more accurate representation of the current rural situation. The proportion sampled in the 15–19-year-old age group was slightly higher. Otherwise, the sample age structure was very similar to the rural population age structure in Assiut. The rural population age structure for Kafr El Sheikh, also based on the CAPMAS 1986 census data, is shown in Figure 3 along with the age structure of the sample for Kafr El Sheikh. Again, a similar pattern of population and sample age structure can be seen. Figures 2 and 3 are given as typical examples and in all governorates, the same pattern in the 0–4-year-old age group was seen. The older age groups overlapped more closely. The comparison of rural population age structure to sample age structure was made for all EPI 1, 2, 3 governorates as an approach to evaluating the representativeness of the sample. The similar pattern of results observed in all 9 governorates suggested a good representation of age in the sample.

After the data from each governorate were collected, the probability of selection was calculated for each record or individual and weighted. The sum of the weighted sample was examined and compared with the respective rural population census figures. In all but 2 cases, the sum of the individual sample weights agreed with the rural population (after making age specific adjustments of the weights for non-response). This ensured that each individual selected was accounted for, the correct probability and weight was established, and that reliable estimation of parameters can be made.

As mentioned earlier, this assessment could not be completed for 2 governorates (Minufiya and Ismailia) due to
problems in the implementation of the selection of ezbas and households in ezbas by the respective field teams. Accordingly, sampling weights could not be calculated as the probability of selection in each of these governorates. The calculation of SEs are also dependent on a known probability of selection at each stage and could not be determined for these 2 governorates. Therefore, the parameter estimates (overall prevalence and GMEC) cannot be known, although the patterns of infection and associations of infection or morbidity with various factors may still be meaningful. This limitation should be considered when interpreting the results from Minufiya and Ismailia.

The relative complexity of the EPI 1, 2, 3 sample design was due in part to complex and multiple objectives. However, a major factor was the complex relationship of village administrative centers and their associated ezbas for which there was no readily available sampling frame. This would not be a design issue for studies that do not need ezbas.

Large studies in Egypt have used sample designs that were stratified by districts within the governorate, a feature not included in our design. Unfortunately, these studies did not include the district level population weight and variance into consideration at the analysis stage. This produces biased estimates of prevalence. We found that developing a sample design that would provide good district specific estimates, where the second stage sample is usually villages, may require an excessively large number of sample villages per district. This and other issues justifies a careful evaluation of the objectives of the project before finalizing a sample design. In large studies, the evaluation of the objectives and the interrelated impact on the sample design and sample size cannot be over-emphasized.

The reports that follow indicate that the most significant accomplishments of the EPI 1, 2, 3 sample design were the precise prevalence estimates, intensity of infection, and other parameters. The SE for overall estimates, village- and ezba-specific estimates, and age- and sex-specific estimates were well within acceptable ranges for purposes of fulfilling the stated objectives of the project.

Nevertheless, there are other limitations of this sample design that need to be stated. This sample design allows for comparison of results within each governorate and between the selected governorates. However, no inferences can be drawn or extrapolated to governorates that were not selected. Accordingly, overall estimates for Egypt have not been and cannot be made. The reader should not make inferences about other governorates not included in the sample. Moreover, caution is advised in comparing the estimates from this study to other governorate-wide studies, which have not used sampling designs proportional to the population or have not estimated SEs. Finally, data from 1 of the selected governorates should not be combined with that from other governorates for the purposes of increasing the sample size or to get combined estimates for all of Upper Egypt or the Nile Delta. Overall estimates from data combined in this fashion will be biased.

Financial support: This research was supported by the Egyptian Ministry of Health/United States Agency for International Development-funded Schistosomiasis Research Project, 263-0140.2.

Authors’ addresses: Mohamed H. Hussein, Department of Community Medicine, Kasr El-Aini Faculty of Medicine, University of Cairo, Cairo, Egypt. Medhat K. El-Sayed, Pharmacology Department, Faculty of Medicine, Suez Canal University, Ismailia, Egypt. Maha Talaat, Community Medicine Department, Theodor Bilharz Research Institute, Cairo, Egypt. Amal El-Badawy, Community Medicine Department, Faculty of Medicine, Zagazig University, Zagazig, Egypt. F. DeWolfe Miller, School of Public Health, University of Hawaii, Honolulu, HI 96822.

Reprint requests: Schistosomiasis Research Project, Medical Service Corporation International (MSCI)/SRP, 1716 Wilson Boulevard, Arlington, VA 22209.

REFERENCES