PN-AAZ-406

Sample Design
Analytical Plan
and
Data Processing Recommendations
for the
Gezira-Managil Area Monitoring Study
of the
Blue Nile Health Project

Karen Stanecki, Mathematical Statistician
Steve Tourkin, Survey Statistician
Jeff Rosen, Systems Analyst
International Statistical Programs Center
U.S. Bureau of the Census

July 1984

Table of Contents

Introduction

I. Study Background and Objectives

II. Program Monitoring and Collection of Village Characteristics

III. Design of the School Children Survey

- A. Sampling Frame
- B. Sample Selection
- C. Sampling Weights
- D. Sample Design Implications and Recommendations
- E. Child Questionnaire
- F. Record Layout
- G. Specification for Recodes

IV. Analytical Plan

- A. Short-term Effects of the Program
- B. Long-term Effects of the Program
- C. Further Detailed Research
- D. Statistical Testing

V. Data Processing Considerations

- A. Data Entry and Editing
- B. Data File Creation and Recoding
- C. Generation of Tables and Reports

VI. Software Requirements

VII. Hardware Recommendations

BLUE NILE HEALTH PROJECT
GEZIRA-MANAGIL AREA MONITORING STUDY

Introduction

This report documents the technical assistance provided to the Blue Nile Health Project staff of the Ministry of Health in Sudan by the U.S. Bureau of the Census at the request of the U.S. Agency for International Development Mission in Khartoum. Two statisticians, Karen Stanecki, and Steve Tourkin, from the U.S. Bureau of the Census travelled to Khartoum to work with the Research and Training Unit of the Blue Nile Health Project.

A basic study design had been established to monitor the effects of the chemotherapy and snail control measures in the Gezira-Managil Area of Sudan. The U.S. Bureau of the Census was requested to assist in developing a monitoring and evaluation system. This included designing the sample for selecting schools into the study, designing the questionnaire and other information-gathering forms and providing suggestions for the data processing hardware and software requirements. Based upon the proposed study design, a sampling plan was developed, a questionnaire was designed and pretested in the field, data processing specifications were developed and hardware and software considerations were discussed.

I. Study Background and Objectives

Past studies in the two million-acre Gezira-Managil irrigated area in Sudan by the Blue Nile Health Project Staff in the Ministry of Health and the Gezira Schistosomiasis project have found a prevalence rate for schistosomiasis of over 50 percent, with approximately 80 percent of school-aged children infected. The Blue Nile Health Project is a program of mass chemotherapy against schistosomiasis in 70 percent of the land area. As chemotherapy takes place, a complementary program of recurrent focal mollusciciding to control the snail population is carried out. Further transmission control is expected by health education programs implemented by specially formed village committees, by the improvement of village water supplies and by the control of aquatic vegetation which at present is blocking the minor canal network of the irrigation system.

The overall objective of the study proposed by the Research and Training Unit in the Ministry of Health is to monitor the Blue Nile Health Project's large-scale schistosomiasis control program in the Gezira-Managil irrigated area of Sudan.

The specific objectives of the monitoring study are:

(1) to collect trend data on the prevalence, incidence, and intensity of schistosomiasis throughout the Gezira-Managil irrigated area from school children and residents living in agricultural communities, and

(2) to monitor the efficacy of a control campaign based on mass chemotherapy and transmission control.

It is felt that the most reliable indicators of incidence of schistosomiasis will be obtained by examining school-aged children (6-16 years), the group with the highest prevalence. School children are the most feasible group from which repeated stool and urine samples can be collected. The Blue Nile Project Research and Training Unit, therefore, will depend heavily upon these children as the major source for clinical data.

The Gezira-Managil region is divided into six areas according to the schedule for intervention by chemotherapy and molluscicide applications. The six areas are made up of groups of administrative agricultural subunits called blocks, each containing a block office and an agricultural inspector. There are 102 blocks in the Gezira-Managil.

The total number of schools to be selected from each area is limited by the maximum number that can be surveyed in a 1-year period. Due to ethical reasons, it is felt that all children in a selected school should be included in the study, therefore severely limiting the number of schools to be selected. All the schools in the sample will be visited (school children interviewed and clinically tested) every year.

In areas 1 and 2, where chemotherapy and mollusciciding have already taken place, eight primary schools will be selected. Every pupil in each of the schools will be examined annually. This will provide the data for the majority of the treatment assessment.

In areas 3, 4, and 5, where treatment will be done in years 3, 4, and 5 of the project, six schools will be selected to provide a cohort of children to be surveyed annually. In the first year of the study, only children in classes one and two will be interviewed. At each subsequent visit, the new students in each school will be added to the cohort.

Area 6, which will be the last area to be treated, will provide a control[1] group to compare with areas 1 and 2. Eight schools will be selected and the total school population examined annually. However, the main interest in area 6 is to follow the natural history of the disease by studying the clinical data.

During the control program, information will be obtained on the program and specific village characteristics that will provide the basis for assessing the success of the snail control (focal mollusciciding) and the mass chemotherapy.

---

[1]Areas 1 and 2 are different from area 6 and therefore cannot be a true control area. The areas differ from one another in that area 6 is the furtherest away from easy access whereas area 1 is the easiest accessible to the project (nearest the main towns and within close travel distance).

II. Program Monitoring and Collection of Village Characteristics

In addition to the data being obtained through a survey of school children, data will be collected from the villages where the school children live. This data will come from the various other units of the Blue Nile Health Project and will be coordinated by the Research and Training Unit. The data will be used in analyzing the data collected in the survey of school children. (See section IV.C).

The data collected will include the following components:

1) reports from the Epidemiology Unit of the Blue Nile Health Project regarding the chemotherapy program (this includes a village census conducted at the time of mass chemotherapy);

2) reports from the Biology Unit regarding details of the snail control measures (when and how often focal mollusciciding is done);

3) reports from the Community Participation Unit regarding the health committees and health education;

4) results of quality control observations on the snail control performed by the Research and Training Unit; and

5) information provided by school headmasters during the yearly survey of school children (for example, reasons for absences of children).

The data will be transferred to a "Village File" that will be kept in a folder for each village along with all forms and questionnaires for that village. The proposed data items that will be contained in this file are shown on the attached forms.

III. Design of the School Children Survey

A. Sampling Frame

The Blue Nile Health Project has a complete list of villages with schools by block (not all villages contain schools). Villages have either separate boys' and girls' schools or a "mixed" school. In those selected villages that have separate schools, both will be included.

In order to reduce fieldwork costs, it was decided to select approximately 60 percent of the schools from villages that have mixed schools. Five of the eight schools in areas 1, 2, and 6 and four of six schools in areas 3, 4, and 5 will be selected from villages with mixed schools (Table 1). Therefore, the list of villages with schools will need to be stratified into villages with mixed schools and villages with separate schools for each of the 6 areas.

The villages from each area for each school type should be listed in a geographical fashion, i.e., those villages near each other should be listed consecutively. The villages should be numbered on each list from 1,2,...,S. From this geograpically ordered and numbered list, schools can be systematically selected with a random start.

Table 1. Number of Villages to be Selected by Area and Type of School*

Areas	Villages with Mixed Schools	Villages with Separate Boy and Girl Schools
1, 2, and 6	5	3
3, 4, and 5	4	2

*There will be a separate list of villages for each area by the type of school(s) in the village.

B. Sample Selection

The number of schools, n, to be selected from a given list (for example five villages will be selected from the list of villages with mixed schools in area 1) will be used to determine the interval, I, for selection. The total number of villages on the list, S, is to be divided by n, the number of schools to be selected. The result (rounded to the nearest whole number) will be the interval number I. A random number, r, between 1 and I is then selected. This will correspond to the village number on the list and will be the first village selected. The sum of r+I will give the corresponding number of the second village to be selected. Continuing to add I to the previous sum in this way will designate the rest of the villages to be selected from the numbered list of villages.

C. Sampling Weights

The estimation procedures will involve applying a weight (an expansion factor) to computer record for each sample child equal to the inverse of the probability of selection. The weights will vary by area and school type so there will be at least 12 different weights. The weight can vary by village if there is more than one school that can be selected.

The weight for a sample child from a mixed school in village j in area i can be expressed as:

$$W_{ij1} = \frac{N_{i\cdot 1}}{n_{i\cdot 1}} \cdot \frac{M_{ij1}}{m_{ij1}}$$ [2]

where

$N_{i\cdot 1}$ = total number of villages with mixed schools in area i

$n_{i\cdot 1}$ = 5, number of villages with mixed schools selected where i=1,2, and 6

= 4, number of villages with mixed schools selected where i=3,4, and 5

M_{ij1} = total number of schools in village j in area i

m_{ij1} = 1

Subscripts:

i = treatment area 1,...,6

j = number of villages selected

1 → villages with mixed school

The weight for the record of a school child from a boy or girl school can be expressed as

$$W_{ij2} = \frac{N_{i\cdot 2}}{n_{i\cdot 2}} \cdot \frac{M_{ij2}}{m_{ij2}}$$ [2]

where

$N_{i\cdot 2}$ = total number of villages with separate boy and girl schools in area i

---

[2] In most cases the fraction will be equal to one.

$n_{i \cdot 2}$ = 3, number of villages with separate boy and girl schools selected where i=1,2, and 6

= 2, number of villages with separate boy and girl schools selected where i=3,4, and 5

M_{ij2} = total number of schools in village j in area i, i.e., the sum of both boy and girl schools in the village (usually 2)

m_{ij2} = 2, number of schools selected in village j in area i

Subscripts:

i = treatment areas 1,...,6

j = village selected

2 → villages with separate boy and girl schools

D. Sample Design Implications and Recommendations

A more efficient sample design could be achieved by selecting more schools and subsampling students within schools. However, the project staff feels that it must include all the students in a selected school. This severely limits the number of schools the team can visit within a year.

The data resulting from such a highly clustered sample (clusters of approximately 300 school children) will result in correlated data. That is, the data from individual school children in a selected school will be very similar and as a result, the reliability of the resulting estimates will be adversely affected. It should also be noted that computer package statistical programs produce variance estimates that are based on the assumption of simple random sampling. In this case, the statistics produced would give a very liberal estimate of the variation. Therefore, it is essential that a special program be developed which will calculate variance estimates based on the sample design of this survey. Variance estimates will be required for the analysis (see section IV.D).

E. Child Questionnaire

The questionnaire to be administered to the children in the sample is attached. It is designed to obtain annual data on treatment for schistosomiasis, household water source and presence of latrines and to record symptoms evident during an annual physical examination and laboratory results from tests performed after the examination. The questionnaire also includes identification information and demographic characteristics. The questionnaire is formatted so that all questions are on the left side of the page and all answer categories to be keyed into the computer record are on the right side. Each answer that is to be keyed has a circled field identification number next to the answer space.

F. Record Layout

A record layout is also attached. It displays the field identification and character position on the computer file for each variable collected on the questionnaire and additional variables which are recoded or constructed from the questionnaire variables. Also shown are the valid values for each variable. The record layout will be used in developing specifications for analytic tabulations.

G. Specifications for Recodes

Variables 1 through 41 (referred to as C1, C2, ... C41) are keyed directly from the questionnaire. Variables 42 through 54[3] will be created by a recode program. The specifications for developing these variables (to the extent possible at this time) are listed below.

1. Area Number - C42

   This will be based on C2, block number. All block numbers will be recoded into values 1, 2, 3, 4, 5, or 6 for area. The blocks included in each area are not known at this time.

2. Village Type - C43, C44, C45, C46, C47

   These will be based on C12, village code. The specific groupings and values for these cells will not be established until later in the evaluation. Examples of these cells are discussed in section IV.

3. Sample Weight - C48

   This will be based on the formulas in section III.C.

4. Change in Egg Count - C49

   This recode will be based on C36 and C39. It will be computed by the following formula:

   (C36 + C39)a - (C36 + C39)b

   where a and b are the years of interest (identified in C7 - year, or C1 - visit number). The values of "a" should be from the more recent year.

5. If Ever Treated by Chemotherapy - C50

   This will be based on C15. If C15 = 1 for any year, C50 = 1. If C15 = 2 for all years, C50 = 2.

---

[3]This includes the recoded variables known at this time. Others will be developed later beginning with C55.

6. If Positive for Schistosomiasis - C51

This will be based on C36 and C39. If C36 <u>and</u> C39 both = 0, C51 = 2. If C36 <u>or</u> C39 > 0, C51 = 1.

7. Frequency of Treatment - C52

This will be based on the values in C15. If C15 = 2 for all years, C52 = 1. If C15 = 1 for one year and 2 for all other years, C52 = 2. If C15 = 1 for more than one year, C52 = 3.

8. History of Prevalence - C53

This will be based on the recoded values in C51 for two specific years. It will show, for those two years, whether the child was negative for schistosomiasis in both years, changed from positive to negative or changed from negative to positive. Since the years of interest might vary by area, a separate set of specifications for each area (C42) may be necessary.

Values of C51

Earlier Year	Later Year	C53
2	2	1
1	2	2
2	1	3

9. Total Egg Count - C54

This is the sum of C36 and C39.

IV. Analytical Plan

The analysis will address several issues - the short-term effects of the program, the longer term effects of the program, an explanation of why observed changes have or have not occurred, and additional health related research. The first three of these are discussed separately below. The tabulations that will be produced in order to make these assessments are expressed in terms of the "field ID's" of the variables contained in the record layout (attached).

A. Short-term Effects of the Program

The most important tabulations to examine in order to assess the short-term effects of the program will be the overall prevalence rates and mean egg counts for the villages and the areas. For areas, this will involve the cross tabulation of C51, C36 and C39 by C42.[4] Differences in the measurement between year 1 and year 2 will show the short-term trend. For example:

Exhibit 1. Prevalence Rates and Mean Egg Counts by Area, 1984 and 1985 (Illustrative Values Only)

Area	C51 Prevalence Rates (Percent)		C36 Mean Egg Count in Stool		C39 Mean Egg Count in Urine	
	1984	1985	1984	1985	1984	1985
All Areas	60	54	34	24	2	1
I	62	38	20	12	1	1
II	58	59	60	62	1	1
III	60	59	19	20	16	12
IV	54	52	5	3	5	4
V	61	64	23	24	3	1
VI	66	66	28	30	2	1

The tables showing the rates and egg counts per village will be produced by cross tabulating cells 51, 36 and 39 by C12.

---

[4] Note that this tabulation will require computation of the mean for cells 36 and 39. An alternative table could show categories of egg counts (e.g., 0-20, 21-40, 41 or more). This would require either recoding cells 36 and 39 into separate cells or additional tabulation programming.

In the short term, only some of the villages will have received treatment. Additional tables will look at the prevalence rate and egg counts for treated versus nontreated villages (cells 51, 36, 39 by one of cells 43 to 47[5]). The treated villages can then be further examined by cross tabulating cells 51, 36 and 39 by the relative success of the project intervention in terms of snail control and chemotherapy compliance (two of cells 43-47).

Further investigation beyond village or area level estimates of the health indicators can proceed by using the results for each child. At this stage it would probably be valuable to cross tabulate the incidence of schistosomiasis (C51) and the mean egg count (cells 36 and 39) by whether the child was actually treated by chemotherapy (C15) and by whether the treatment was taken within the previous 6 months (C16). It might also be useful to know which children have been treated, since it is possible that compliance will differ among various groups. This would be accomplished by cross tabulating whether the child was treated (C15) and whether the treatment was project or nonproject (C18) by ethnic group (C14), sex (C11) and age (C13). A table of C19 will show the reasons for not receiving treatment.

The tables produced at this point should describe the overall levels of the health indicators for villages and areas, how these rates differ between villages which have or have not received treatments, and the association between incidence and treatment and by the relative success of the treatment, and who has received the treatment in individual children. Further examination of the data to explain why children remain "positive" for schistosomiasis in treated areas is addressed section C. One last group of tables that might be interesting for the short-term analysis would be the symptoms that the positive children have by prevalence rate (and perhaps by mean egg count as well). These would be cells 24 through 35 by C51, C36 and C39.

---

[5]These will be coded onto the file based on the village files. Additional groupings of "village type" in terms of other characteristics can be coded into other fields as well. See section C for details.

B. Long-term Effects of the Program

The key tables that will be used to assess the effects of the program after 4 or 5 years will again be the prevalance and mean egg counts by area and village. For example:

Exhibit 2. Prevalence Rate by Area, 1984-1988
(Illustration Only)

Area	1984	1985	1986	1987	1988
All Areas	60	54	12	15	22
I	62	38	6	9	12
II	58	59	10	10	10
III	60	59	2	12	11
IV	54	52	20	20	25
V	61	64	4	6	6
VI	66	66	58	60	56

This table along with the mean egg count tables would show the trend in prevalence and are the major indicators of project success. The mean egg count trend will show the changes for the entire group of children. It will be interesting to also look at the change in mean egg count per child. That is, to examine to what degree intensely infected versus mildly infected children have experienced a decrease in egg count. This would be accomplished by cross-tabulating the recoded variable "change in mean egg count" (C49) by categories of the total egg count (C54) in the year before treatment. For example:

Exhibit 3. Change in Egg County by Intensity Before Treatment

Egg Count Year Before Area Treated (1984)	Change in Egg Count, 1984-1988			
	1,000 or more	100-999	10-99	0-9
All Children				
1,000 or more				
100-999				
10-99				
0-9				

Another useful cross-tabulation would show the rate of reinfection. The prevalence rates will give an overall measure of reinfection, but it is

necessary to examine the data for each child to actually determine how many children change from negative to positive. This table would involve a recode of C51 for the years of interest.[6] For example:

History of Prevalence, Area I, 1985-6

Result	Number of Children	Percent
Negative Both Years	XXXX	XX
Positive to Negative	XXXX	XX
Negative to Positive	XXXX	XX

C. Further Detailed Research

If the summary tables described in sections A and B show that the prevalence rates have dramatically decreased, and that few children are tested as positive for schistosomiasis by the end of the project, it is probably not necessary to conduct further research. If many children do remain positive, however, it will be desirable to examine (to the extent possible) the reasons.

This objective will be accomplished by observing associations between the main dependent variables, prevalence and mean egg count, and other variables collected. The analysis will be divided into two parts. The first part, a "bivariate analysis," examines associations between pairs of variables, for example, the association between prevalence and success of snail control. In the bivariate part of this analysis, the question examined is whether two variables are associated in a statistically significant manner; that is, whether the observed differences in the distribution of one variable at the different levels of the other variable are beyond the range that could be attributed to sampling variability alone. The second part, a "multivariate analysis," expresses more complex relationships that may arise among three or more variables. The multivariate analysis extends the bivariate analysis by considering the joint effect of the variables on each other. This general notion may be illustrated by the following example. Suppose that prevalence was shown in the bivariate analysis to be related to both snail control and household water source separately. Multivariate analysis of the three variables simultaneously could produce any of the following outcomes:

a. Prevalence is associated with snail control, but water source has no additional effect when snail control is considered.

b. Prevalence is associated with water source, but snail control has no additional effect when water source is considered.

---

[6] This recode will differ for different areas, since analysts will want to look at reinfection rates in the year after treatment and subsequent years. The year of treatment will vary by area.

c. Both snail control and water source have effects on prevalence. The separate effects of snail control and water source combine in an additive manner; i.e., the effect of snail control does not depend on the effect of water source, and visa versa.

d. Both snail control and water source have effects on prevalence, but the effect of one variable depends on the level of the other.

In the first two outcomes, one of the two variables "accounts for" or "explains" the effect of the other variable on prevalence observed in the bivariate analysis. In the last two outcomes, both variables affect the prevalence with neither fully explaining the effect of the other. In the third outcome, the effect of each variable on prevalence does not depend upon the level of the other variable, whereas in the last outcome, both variables may be said to interact in their effect. The distinction between three and four requires statistics beyond the scope of the proposed analysis. It will be possible to recognize that interactions exist but not to quantify the separate effects due to each variable.

1. Bivariate Analysis

There are two groups of variables that should be cross-tabulated with prevalence - those originating from the village file and those originating from the child questionnaire. The village file contains information on the snail control, chemotherapy program, village water and sanitation, health education program, and several other variables that will provide a comprehensive profile of each village during the study period. The analysts will "group" the villages in several ways to differentiate between relative differences in the above measures. Using the village code (C12), a program can be written to enter numerical values for these created variables on the child file.

Fields 43 through 47 are reserved for this purpose, and additional fields can be utilized if more than five groupings are desired. Examples of variables that could be created from the village file are:

a) Success of intervention

1 - satisfactory (could be more than two categories)
2 - less satisfactory

b) Success of intervention and other factors

1 - intervention satisfactory; no other changes
2 - intervention satisfactory and improved water system
3 - intervention satisfactory and improved latrines
4 - intervention satisfactory and better health practices
5 - intervention satisfactory and better water system and latrines
6 - intervention satisfactory and better water system and health practices
7 - intervention satisfactory and better water system, health practices and latrines
8 - intervention less satisfactory; no other changes
9 - intervention less satisfactory and improved water system
10 - intervention less satisfactory and improved latrines
11 - intervention less satisfactory and better health practices
12 - intervention less satisfactory and better water system and latrines
13 - intervention less satisfactory and better water system and health practices
14 - intervention less satisfactory and better water system, health practices and latrines

The created variables should be based on the observed differences between the villages. The established categories can attempt to distinguish between several characteristics at the same time (number 2 above) or focus on one characteristic (number 1 above). Another possible recode would involve assigning numerical scores to each of the major characteristics on the village file and summing these to create an overall score describing the village. A recode could then group the villages based on this number.

The second group of variables for the bivariate analysis will be taken from the child file directly. These include:

C11 - sex
C13 - age
C14 - ethnic group
C15 - whether child was treated
C20 - household water source
C21 - presence of latrine

As mentioned, the bivariate analysis will show which of these variables relate to prevalence. The variables having the strongest relationships should be further examined in the multivariate analysis.

2. Multivariate Analysis

The variables that will be examined together in this analysis will depend on the results from the bivariate analysis. The strategy will involve producing 3-way cross-tabulations that look at prevalence rates in terms of one variable while controlling for the effects of a second variable. This can be best illustrated by the following example.

Suppose that the bivariate analysis were to produce the following results.

Exhibit 4.

	Percent Positive
Success of intervention	
1. satisfactory	10
2. less satisfactory	30
Household Water Source	
1. piped	22
2. wells	23
3. canals	25
Presence of latrines	
1. yes	8
2. no	35
Ethnic Group	
1. Arab	17
2. Western Sudanese	23
3. Nigerian	26
4. Other	28

In this example, the variables that appear to influence the prevalence rate to the greatest extent are the intervention and the presence of latrines.[7] The water system and the ethnic group do not appear to have much effect on prevalence.

In order to examine the relative effects of intervention versus latrines, a 3-way tabulation showing the prevalence rates by both these variables would be produced. For example:

Exhibit 5. Prevalence Rates by Success of Intervention of Presence of Latrines

Level	Total	Latrines	No Latrines
Intervention Satisfactory	10	7	12
Intervention Less Satisfactory	30	28	39

This table would suggest that the predominant difference in prevalence was due to the intervention and is affected only slightly by the presence of latrines.

---

[7]The level of significance of the difference in prevalence will be established by conducting t-tests.

The 3-way tabulations should indicate which variables have the greater influence on prevalence but will not quantify the relative difference. It would be possible to accomplish this if additional programs are developed that enable statistical modeling.

D. Statistical Testing

A statistical test that is used to test the significance of the observed differences in sample data is the t-test. When the statement is made that difference is statistically significant, it is meant that there is only a small probability that the observed difference is an accident of sampling. The estimates are subject to error arising from the fact that they were obtained from a sample survey rather than from a complete census. The particular sample used is one of a large number of possible samples of equal size that could have been used applying the same sample design and selection procedures. Estimates derived from different samples would differ from each other. The standard error of a survey estimate is a measure of the variation among the estimates from all possible samples and is, therefore, a measure of the precision with which the estimates from a particular sample approximates the average result of all possible samples. As was mentioned earlier, a customized computer program should be written to calculate the standard errors of the estimates for the analysis.

The estimate and its associated standard error may be used to construct a confidence interval having a prescribed probability that would include the average result of all possible samples. The chances are 68 out of 100 that the survey estimate would differ from the average results of all possible samples by less than one standard error. Similarly the chances are about 90 out of 100 that the differences would be less than 1.6 times the standard error, about 95 out of 100 that the differences would be less than 2.0 times the standard error; and about 99 out of 100 that the differences would be less than 2.5 times the standard error.

In using the t-test to test the significance of the difference between two survey estimates, one needs the standard error of the difference. The standard error of the difference is approximately equal to the square root of the sum of squares of the standard errors of each estimate considered separately. If the standard error of the difference is less than the difference itself, the difference between the two estimates is statistically significant at the 68 percent confidence level; if twice the standard error of the difference is less than the difference, then the difference is significant at the 95 percent confidence level.

For example, suppose the mean egg count for area one in 1984 was $\bar{x}_1$ and in 1985 it was $\bar{x}_2$ with associated standard errors of σ_1 and σ_2 respectively. We could test for significant difference by first calculating the standard error of the difference σ_Δ.

$$\sigma_\Delta = \sqrt{\sigma_1^2 + \sigma_2^2}$$

To test for the significance of the difference in mean egg counts between 1984 and 1985 at the 95 percent confidence level, one would compare

$$\bar{x}_1 - \bar{x}_2 \text{ to } 2\sigma_\Delta$$

If $\bar{x}_1 - \bar{x}_2 > 2\sigma_\Delta$ we would be able to say that there is a significant difference at the 95 percent level of confidence in the mean egg count between 1984 and 1985.

V. Data Processing Considerations

This section will address those computer software and hardware issues which should be considered in developing the microcomputer-based data processing system for the Blue Nile Health Project study. A discussion of the basic system design in this section will be followed by specific software requirements and microcomputer hardware recommendations in Sections VI and VII.

The data processing system should consist of three major components: (A) data entry and editing of the survey questionnaires, (B) creation and recoding of the data file(s) and (C) generation of tables and reports.

A. Data Entry and Editing

Both the data entry and editing components are considered in the same subsystem. This is done for several reasons. One can use existing microcomputer software to format the microcomputer display screen to resemble the child questionnaire form. Thus, when the data are entered, keying errors will be reduced. Furthermore, this same data entry software can also be used to check the keyed data for reasonableness. For example, if the data entry clerk keys "3" for a data item response with valid values of "1" or "2," the software will alert the keyer to the mistake. More importantly, the system could prohibit the keyer from proceeding to key the next data field until the existing field is correctly entered. This technique, known as "intelligent data entry" is of extreme importance in building accurate data files. Some intelligent data entry systems even permit cross-checking one or more data items with each other to guarantee consistency. For example, if question 18 can be skipped based on the answer to question 17, then a good intelligent data entry system should be able to guarantee this. If a consistency error should occur that is not attributable to data entry, but rather to an incorrectly completed questionnaire, then a subject-matter specialist could look at the questionnaire and inform the keyer of the correct responses.

B. Data File Creation and Recoding

After the survey data have been keyed and edited, each child record should be organized logically (sorted) by its unique identification code. This identification code, or key, would consist of the school number (C4), the child number (C5) and the year of visit (C1). Assuming that the child questionnaires will arrive for data entry on a flow basis (that is, questionnaires for some villages will arrive in Khartoum after those from other villages), the newly keyed questionnaire records must be inserted into the data file according to their keys. Thus, it will be necessary to have some mechanism to insert them.

To facilitate this and other data manipulation, it would be helpful to think of each year's survey data as being organized into a relational database. In a relational database, questionnaire records comprise the rows of a matrix. For example:

Year 1:

Key:	
Record 1	All Data Items for the First Record
Record 2	All Data Items for the Second Record
	⋮
Record n	All Data Items for the Last Record

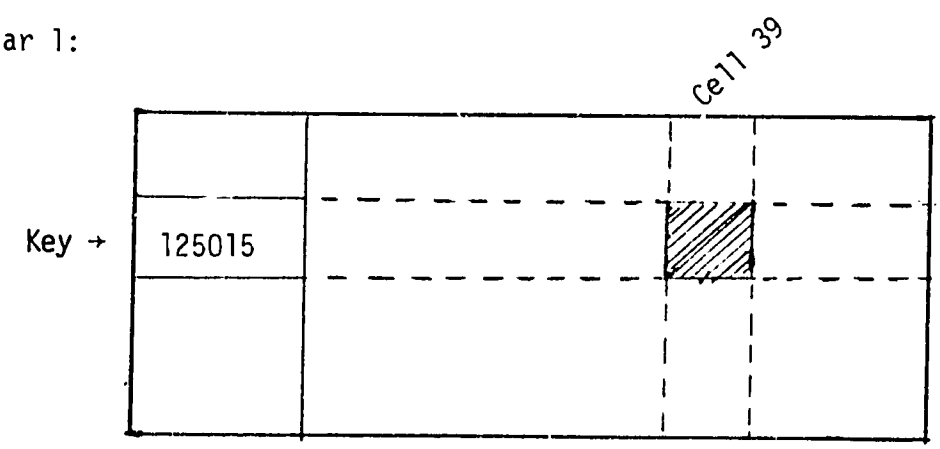
The data items, or responses, comprise the columns:

Year 1:

Cell 1	Cell 2	. . . . .	Cell m
All responses to C1	All responses to C2	. . . . .	All responses to last cell

The intersection of the rows and columns would represent a given data item for a particular questionnaire. For example, the egg count (C39) for a certain child (C05=015) in a certain school (C04=125) will be referenced as follows:

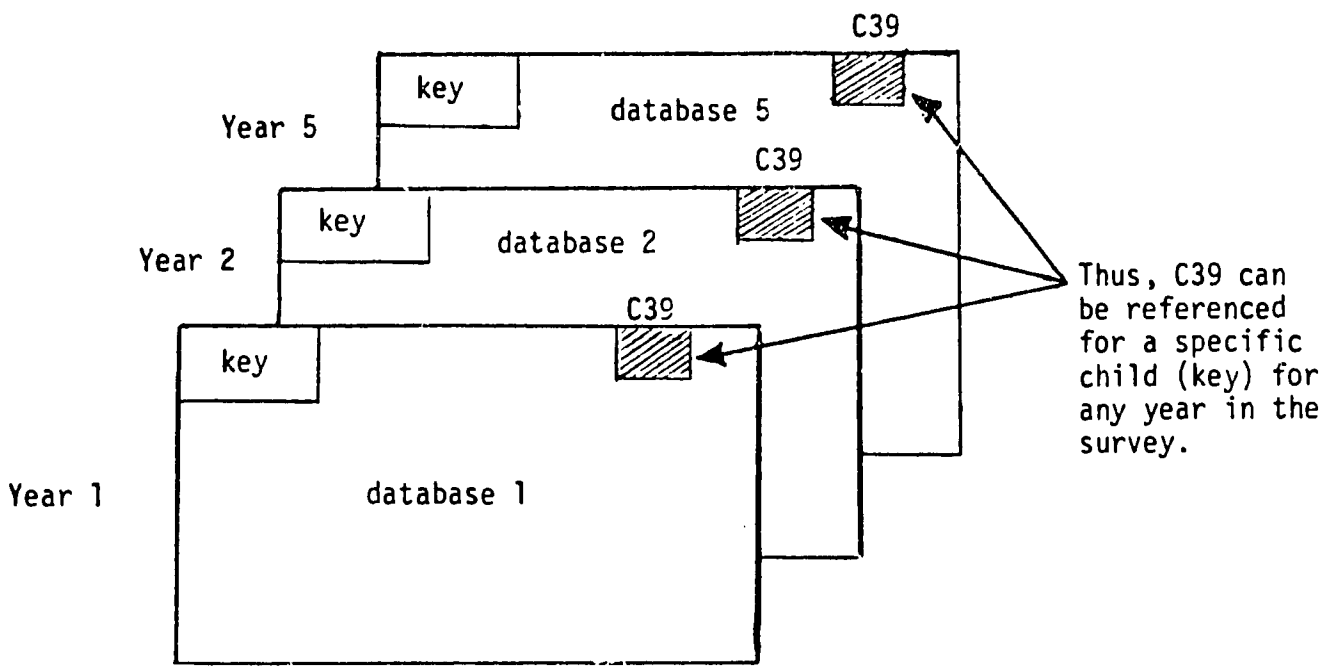
Year 1:



One can create cell recodes for each questionnaire as long as the component cells are present in each child record. For example, the village type recode (C43) is based on C12, the village code, and is easily created for each child record. However, the recode C49 (change in egg count) is determined by comparing two different years' egg counts:

$$(C36 + C39)_{\text{year b}} - (C36 + C39)_{\text{year a}}$$

Therefore, the two databases for year a and year b must be accessed at the same time, with the resultant recode placed on the year b database record. To accomplish this and other related tasks, it would be helpful to picture the five databases for the survey as follows:



C. Generation of Tables and Reports

Given that the data are organized into relational databases as explained previously, it would be easy to generate listings or reports using the database management system (DBMS) language commands. There would be no need to write customized programs to access the data, since the DBMS commands would suffice. For example, a listing of selected characteristics of children whose egg counts (C36 + C39) are greater than a specified amount could be produced for different areas. Furthermore, simple cross-tabulations could be produced using predefined sets of the DBMS commands, referred to as procedures. Tabulation of the treatment of the child (C15) by ethnic group (C14) could be produced by first using the DBMS command to sort the database on C14 and then aggregate the data using the DBMS command to tabulate the treatment (C15). For the more involved cross-tabulations, it would be necessary to have a computer programmer develop the procedures.

If subject-matter specialists (or end users) want to create graphs of selected data items, a three-step process could be used. This would entail first creating a subfile of the desired data by using the necessary DBMS commands and then converting the data from the DBMS format to external format (accessible by other software packages). This newly created data file would then be used as input to a graphics package, which would create the graphic representation (bar chart, pie chart, etc.) requested by the end user. If more sophisticated statistical analyses were desired, a package like SPSS for microcomputers could be employed.

VI. Software Requirements

There are several DBMS packages that could satisfy the software requirements to one degree or another. A thorough evaluation of the top "contenders" (such as dBASE III, Microrim's R:BASE 4000, Knowledgeman, CONDOR 3, and PC FOCUS) should be made by a qualified data processor. As part of this evaluation, prototype databases should be created and each DBMS package tested to see which of them can feasibly meet the following requirements:

1. Run under PC-DOS operating system.

2. Access at least 6 database files concurrently.

3. Handle database records of up to 200 bytes.

4. Facilitate data entry and editing:

   a. Design screen forms.
   b. Verify entered data by using range checking and/or table look-up.
   c. Correct errors immediately and resubmit.
   d. Create customized error messages for display.

5. Perform ad hoc database queries:

   a. Selectively retrieve certain data records using logic criteria (equal to, greater than, less than or equal to, etc.)

   b. Sort and display the desired fields (cells) from selected records with column headings.

   c. Perform mathematical functions such as totals and averages on selected records.

6. Update the database with additional records, or update specific cells as required.

7. Allow relational access to databases:

   a. Extract and sort fields (cells) from selected records to create a new database.

   b. Match values of common fields from one database to another and create a new database.

   c. Join (or satisfy comparison criteria between) two specified cells from two existing databases to form a new database.

   d. Combine records from two existing databases which have common matching fields.

8. Facilitate cross-tabulation of data.

9. Convert database from DBMS format to a format which will interface with other software packages which have graphics capabilities (such as LOTUS 1-2-3) or statistical packages (like SPSS/PC).

VII. Hardware Recommendations

There are two main reasons that the microcomputer to be used should be an IBM PC. First, good overseas representation for microcomputer equipment support and maintenance is critical. Although IBM does not have in-country representation in Sudan, there is IBM expertise available as close as Cairo, Egypt.

Second, because the IBM PC has become the de facto standard for microcomputers, the majority of new software is developed with this in mind. Thus, when it comes time to choose a database management system (DBMS) package for the data processing system, a larger selection of DBMS software will be available.

Because of the prevalence of dust and other problematic environmental considerations, a sealed hard disk drive should be used to hold the DBMS files and other important software as well. The IBM PC XT is delivered with both a 10-megabyte hard disk and a floppy disk drive installed.

It is highly recommended that the standard RAM (Random Access Memory) be increased to at least 320K. This would permit the use of many software packages that require large amounts of memory. If a significant number of mathematics will be performed (i.e., SPSS), then an 8087 math co-processor might prove useful.

If graphics are desired (instead of monochrome text), a color graphics adapter board will be required. A color monitor (screen) will be needed to display the graphs. However, resolution for text and numbers is clearer on the monochrome monitor. It is possible to have both a color graphics monitor and a monochrome monitor attached to the computer. However, both the monochrome and color graphics adapter boards will be required.

Finally, it is recommended that an investigation be made into the use of a hard disk data file back-up and recovery system to guard against loss of data through hardware or user-caused disasters. There are several back-up systems available, some using magnetic tape cartridges.

June 14, 1984

RECORD LAYOUT FOR CHILD HEALTH

Field ID	Character Positions	Variable Name	Values
1	1	Year of Visit	1-5
2	2-4	Block Number	
3	5	School Type	1 - Elementary 2 - Secondary
4	6-8	School Number	1-
5	9-11	Child Number	1-300
6	12-13	Month of Interview	1-12
7	14-15	Year of Interivew	84-89
8	16	Reason for Absence	1 - Deceased 2 - Temporarily absent 3 - Permanently absent 4 - Other
9	17-36	Child's Name	
10	37	Class	1-9
11	38	Sex	1 - Male 2 - Female
12	39-41	Village Code	1-
13	42-43	Age in Years	6-16
14	44	Ethnic Group	1 - Arab 2 - Western Sudanese 3 - Nigerian 4 - Other
15	45	Whether Child Was Treated	1 - Yes 2 - No
16	46	How Recent Child Was Treated	1 - Within 6 months 2 - More than 6 months
17	47	If Child Took Second Dose	1 - Yes 2 - No 3 - No second dose administered
18	48	If Child Participated in Mass Chemotherapy or Was Treated Some Other Way	1 - Mass Chemotherapy 2 - Other

Field ID	Character Positions	Variable Name	Values
19	49	Why Child Not Treated in Mass Chemotherapy	1 - Absent 2 - Contraindicated 3 - Refused 4 - Other
20	50	Household Water Source	1 - Own Tap 2 - Neighbor's Tap 3 - Village Tap 4 - Well 5 - Canal 6 - Other
21	51	Presence of Latrine	1 - Yes 2 - No
22	52-53	Child's Weight (kms)	15-75
23	54-56	Child's Height (cms)	75-200
24	57	Whether Child Reported Abdominal Pain	1 - Yes 2 - No
25	58	If Child Reported Blood in Urine	1 - Yes 2 - No
26	59	If Child Reported Blood in Stools	1 - Yes 2 - No
27	60	If Child Reported Diarrhea	1 - Yes 2 - No
28	61	If Child Reported Fever	1 - Yes 2 - No
29	62	If Child Reported Other Problem	1 - Yes 2 - No
30	63	If Child Had Enlarged Liver	1 - Yes 2 - No
31	64-65	Liver Size (cms)	
32	66	If Child Had Enlarged Spleen	1 - Yes 2 - No
33	67-68	Spleen Size (cms)	

Field ID	Character Positions	Variable Name	Values
34	69	If Child Had Ascites	1 - Yes 2 - No
35	70	If Child Had Anemia	1 - Yes 2 - No
36	71-75	Egg Count in Stools (per gram)	1-12,000
37	76	Examination: Blood in Stools	1 - Yes 2 - No
38	77	Examination: Mucus in Stools	1 - Yes 2 - No
39	78-82	Egg Count in Urine (per 10 ml.)	1-12,000
40	83	Examination: Blood in Urine	1 - Yes 2 - No
41	84	Clinician Prescribed Chemotherapy	1 - Yes 2 - No
*42	85	Area Number	1-6
*43	86-87	Village Type - First Grouping	1-16 (Specify each code when known)
*44	88	Village Type - Second Grouping	1-8 (Specify each code when known)
*45	89	Village Type - Third Grouping	1-8 (Specify each code when known)
*46	90	Village Type - Fourth Grouping	1-8 (Specify each code when known)
*47	91	Village Type - Fifth Grouping	1-8 (Specify each code when known)
*48	92-96	Sample Weight	
*49	97-101	Change in Egg Count	0-12,000
*50	102	If Ever Treated By Chemotherapy	1 - Yes 2 - No

Field ID	Character Positions	Variable Name	Values
*51	103	If "Positive" for Schistosomiasis	1 - Yes 2 - No
*52	104	Frequency of Treatment	1 - Never 2 - Once 3 - More than Once
*53	105	History of Prevelance	1 - Never positive 2 - Positive to negative 3 - Negative to positive
*54	106-110	Total Egg Count	0-24,000

VILLAGE FILE

Village Name _____ Village Number _____ School Numbers _____

Area Number _____ Block Number _____ Number of Pupils:

First Year
Second Year
Third Year
Fourth Year
Fifth Year

Population: approx. _____ census _____ Date of Census _____

Ethnic Group (%) (1) (2) (3)

Schools	Elementary:	Intermediate:	Secondary:
	Male	Male	Male
	Female	Female	Female
	Male and Female	Male and Female	Male and Female

Number of: Health Facility_____ Dressing station_____ Health Center_____ Hospital_____

Electricity Yes No

Location	Canal Name	Type	Distance(m)

Water Supply	Type	Description	Coverage
First Year			
Second Year			
Third Year			
Fourth Year			
Fifth Year			

Sanitation	Pit Latrine	Coverage
First Year		
Second Year		
Third Year		
Fourth Year		
Fifth Year		

Health Education	Village Committee	Formation Date	Session Number
First Year			
Second Year			
Third Year			
Fourth Year			
Fifth Year			

Chemotherapy	1. Mass treatment	Date Number treated Response Major noncompliance categories	Arab Western Sudanese Nigerian Other

2. Followup treatment

Year	Number Referred	Response
1		
2		
3		
4		
5		

Molluscicides	Number of Sites	Monthly application	spot-checks
Year			
1			
2			
3			
4			
5			

Weed Control	Year	Canal Type Major/Minor	Date	Description
	1			
	2			
	3			
	4			
	5			

Other Control Measures	Type	Date	Description

FILE FOR "CAMPS"

Camp Name Camp Number

Nearest Village Name Number

Population: Approx._____ Date

Ethnic Group	(1)	(2)	(3)	(4)
Location	Canal Name		Canal Type	Distance

Water Supply	Year	Description	Coverage
	1		
	2		
	3		
	4		
	5		

Sanitation	Year	Pit latrines	Communal latrines
	1		
	2		
	3		
	4		
	5		

June 14, 1984

BLUE NILE HEALTH PROJECT EVALUATION: CHILD QUESTIONNAIRE (FINAL VERSION)

School Name: _____

Visit Number	(1) ☐
Block Number	(2) ☐☐☐
School Type	(3) ☐ 1 = elem. 2 = sec.
School Number	(4) ☐☐☐
Child Number	(5) ☐☐☐
Month of Interview	(6) ☐☐
Year of Interview	(7) ☐☐
If Absent, Reason (mark one box)	(8) 1. ☐ Deceased 2. ☐ Temp. Absent 3. ☐ Perm. Absent 4. ☐ Other

1. Name of Child (enter on right)	(9)
2. What class are you in?	(10) ☐ Enter 1 through 9
3. Sex of child	(11) 1. ☐ Male 2. ☐ Female
4. What is the name of the village that you live in? _____	(12) ☐☐☐ Village Code
5. How old are you? If "don't know" enter best estimate.	(13) ☐☐ Years
6. Interviewer - Determine tribe	(14) 1. ☐ Arab 2. ☐ Western Sudanese 3. ☐ Nigerian 4. ☐ Other

7a. FOR FIRST VISIT ASK:
Were you ever treated for Bilharziasis

☐ Yes - Ask 8
☐ No - Skip to check item

7b. FOR LATER VISITS ASK:
Were you treated for Bilharziasis since our visit?

8. When was the last time you were treated? _____

9. Did you take pills or injections?

☐ Pills - Ask 10
☐ Injections - Skip to 12

10. What color were the pills? _____

11. Did you take the pills all in the same day?

☐ Yes
☐ No

Skip to check item

12. How many injections did you have? _____	
CHECK ITEM — INTERVIEWER - CODE WHETHER CHILD WAS TREATED OR NOT BASED ON ANSWERS TO Q 7-12.	(15) 1 [ ] Treated - Continue with 13 2 [ ] Not Treated — Skip to 18 3 [ ] Cannot Determine
13. INTERVIEWER - MARK TIME SINCE TREATMENT BASED ON Q 8	(16) 1 [ ] Within 6 months 2 [ ] More than 6 months
14. INTERVIEWER - MARK IF CHILD TOOK SECOND DOSE (IF INFORMATION AVAILABLE)	(17) 1 [ ] Yes 2 [ ] No 3 [ ] Only one dose given
15. INTERVIEWER - Was the entire village treated since the previous visit? ☐ Yes - ASK 16 ☐ No. Mark box 2	
16. Did you take the pill along with the other people? [ ] Yes - MARK BOX 1, SKIP to 18 [ ] No - MARK BOX 2	(18) 1 [ ] Mass Chemotherapy - Skip to 18 2 [ ] Other - Ask 17
17. Why were you not treated with them?	(19) 1 [ ] Absent 2 [ ] Contraindicated 3 [ ] Refused 4 [ ] Other
18a. Do you have a water tap in your house or compound? [ ] Yes - Mark box 1 on right and skip to 19 [ ] No - Ask 18b.	
18b. From where do you and your family get your water? Mark Box 2, 3, 4 or, 5 or 6	(20) 1 [ ] Own Tap 2 [ ] Neighbor's Tap 3 [ ] Village Tap 4 [ ] Well 5 [ ] Canal 6 [ ] Other
19. Do you have a latrine at home?	(21) 1 [ ] Yes 2 [ ] No
20. INTERVIEWER - ENTER WEIGHT OF CHILD IN WHOLE KILOMETERS	(22) [ ][ ] Kilograms
21. INTERVIEWER - ENTER HEIGHT OF CHILD IN WHOLE CENTIMETERS	(23) [ ][ ][ ] Centimeters

EXAMINATION - QUESTIONS FOR CHILD

22. Do you have any of the following symptoms? (MARK YES OR NO FOR EACH)	
a. Abdominal Pain	(24) 1 [ ] Yes 2 [ ] No
b. Blood in Urine	(25) 1 [ ] Yes 2 [ ] No
c. Blood in Stools	(26) 1 [ ] Yes 2 [ ] No
d. Diarrhea	(27) 1 [ ] Yes 2 [ ] No
e. Fever	(28) 1 [ ] Yes 2 [ ] No
f. Other - Specify: _____	(29) 1 [ ] Yes 2 [ ] No
PHYSICAL EXAMINATION OF CHILD	
23. Enlarged Liver?	(30) 2 [ ] No 1 [ ] Yes - Size in CM
	(31) [ ][ ]
24. Enlarged Spleen?	(32) 2 [ ] No 1 [ ] Yes - Size in CM
	(33) [ ][ ]
25. Ascites?	(34) 1 [ ] Yes 2 [ ] No
26. Anemia?	(35) 1 [ ] Yes 2 [ ] No

LABORATORY RESULTS

27. STOOL EXAMINATION	
a. Egg Count	(36) ☐☐☐☐☐ Eggs per gram
b. Blood in stools?	(37) 1 ☐ Yes 2 ☐ No
c. Mucus in stools?	(38) 1 ☐ Yes 2 ☐ No
28. URINE EXAMINATION	
a. Egg Count	(39) ☐☐☐☐☐ Eggs per 10 ml.
b. Blood in Urine?	(40) 1 ☐ Yes 2 ☐ No
29. Clinicians prescribed anti-bilharziasis chemotherapy?	(41) 1 ☐ Yes 2 ☐ No