Provision of Research Methods Support for Neglected Tropical Diseases' Research Projects at Eastern and Southern African Centre of International Parasite Control, Kenya Medical Research Institute, Nairobi

Paul Murima Ng'ang'a

Thesis submitted in partial fulfillment for the Degree of Master of Science in Research Methods in the Jomo Kenyatta University of Agriculture and Technology

2011

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature: …… Date: ……

Paul Murima Ng'ang'a

This thesis has been submitted for examination with our approval as University Supervisors.

Signature: …… Date: ……

Dr. John M. Kihoro

Jomo Kenyatta University of Agriculture and Technology, Kenya

Signature: …… Date: ……

Dr. Sammy M. Njenga

Kenya Medical Research Institute, Kenya

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the Almighty God for the gift of life and strength.

I owe a debt of gratitude to all members of staff at Eastern and Southern African Centre of International Parasite Control (ESACIPAC), Kenya Medical Research Institute (KEMRI), who constantly offered help and encouragement during the internship making it both fruitful and enjoyable. Special thanks to Mr. Mwobobia, Ms. Judy, Ms. Masaku and Mr. Kanyi, for providing some of the resources, on which this work is based. I am greatly indebted to Centre for Public Health Research (CPHR), KEMRI statisticians, Mr. E. Muniu and Mr. R. Mwangi, for sharing their vast and superb statistical knowledge, skills and ideas.

My sincerest thanks also go to the Centre Director (ESACIPAC), who was also my supervisor, Dr. Sammy M. Njenga, for sparing his time to facilitate my internship and for sharing his data sets. Without you, all this work would not have been possible.

I would especially like to thank my supervisor, Dr. John M. Kihoro, who always offered a hand of friendship and encouragement at every stage of my studies. His insights and mentorship have greatly influenced my understanding of the research methods.

I sincerely thank the facilitators from Statistical services Centre, University of Reading; Mr. R. Coe and Dr. R. Stern, and, also, Dr. D. Stern and Mrs. B. McDermott for their enthusiasm, vision, unfailing support and commitment, and for making the MSc. Research Methods experience one that I can look back upon with pleasure. As scientific leaders in their own right, they provided excellent mentorship, instilling in me the desire to excel and launching me into what promises to be an exciting and fulfilling career.

I am grateful for the financial support I received from Regional Universities Forum for Capacity Building in Agriculture (RUFORUM).

Last but not least, I would like to thank my parents, brothers and friends, more so Kinyua, who have played a major role in shaping my life and future aspirations and to my dearest and beloved wife, Wanjiru C, and my daughter Wanjiru J, for their unfailing love and moral support.

Having enjoyed doing my MSc, I hope to be a real Research Methods professional some day!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AE	Adverse effect
CPHR	Centre for Public Health Research
CI	Confidence Intervals
EDA	Exploratory and descriptive analyses
ESA	Eastern and Southern Africa
ESACIPAC	Eastern and Southern African Centre of International Parasite Control
HIV	Human Immunodeficiency Virus
ITNs	Insecticide-treated mosquito nets
JKUAT	Jomo Kenyatta University of Agriculture and Technology
KEMRI	Kenya Medical Research Institute
MSc	Master of Science
NTDs	Neglected tropical diseases

PPS	Probability Proportional to Size
RM	Research Methods
RUFORUM	Regional Universities Forum for Capacity Building in Agriculture
SPSS	Statistical Package for Social Sciences
SSC	Statistics Service Centre, University of Reading
STH	Soil transmitted helminthiasis

ABSTRACT

The availability of a large variety of elementary and complex statistical methods and the quick pace of change and development in statistics makes most researchers' statistical knowledge insufficient for them to be independent. It is unrealistic to expect researchers to solve the statistically-related challenges that arise in the course of their investigations on their own. Consequently, support from RM specialists is expected to improve the quality of research outputs by minimizing studies with unsound and unreliable outputs.

Reported in this dissertation are the research methodology tasks undertaken at ESACIPAC (KEMRI) during internship period of one year. The tasks included participation in research by offering consultancy services, data management, statistical analysis and reporting.

The consultancies were diverse as they ranged from technical support in design of projects to reporting of research findings. Reported in this document are consultations involving statistical inputs in drafting sections of a proposal, sample size computations for studies, statistical support in reporting of findings and computation of confidence intervals.

Data management was conducted for the study; "*Adverse effects associated with mass drug administration of praziquantel and albendazole …*" Double data entry was done in Excel. The spreadsheets were constructed with measures for minimizing data entry errors while maximizing processing efficiency, that is, frozen top rows, drop down lists and validation

checks. On data entry completion, auditing was done, and then comparisons using Epi Info. Lastly, data were exported to SPSS and coded further in preparation for analysis.

Data were analyzed for the study; "*Factors contributing to re-infection with S. mansoni among primary school children …*" The sample was described using frequencies, rates and proportions for categorical variables while appropriate measures of central tendency and dispersion were used for continous variables. To identify factors predictive of re-infection, crude associations were explored using bivariate analyses and then logistic regression model constructed. Adjusted odds ratios were computed for the significant factors. The final model showed a reasonable fit when assessed for goodness-of-fit using Hosmer-Lemeshow test (p=0.270). Finally, reporting of the findings was done as a collaborative effort between the investigator and the research methods intern.

CHAPTER 1: INTRODUCTION

1.0 Background

Kenya Medical Research Institute (KEMRI) is a state corporation that was established in 1979. It is the national body responsible for carrying out human health research in Kenya. KEMRI is situated in Nairobi, off Mbagathi road, and has eleven centres, ESACIPAC being one of them. Eastern and Southern Africa Centre of International Parasite Control (ESACIPAC) was established in 2000 with the assistance of the Japanese government under the Global Parasite Control Initiative. The mission of ESACIPAC is to undertake human resource development to strengthen research and control programmes on parasitic diseases in the eastern and southern Africa region, covering Kenya, Uganda, Malawi, Zambia, Zimbabwe, Botswana, Zanzibar and Tanzania mainland.

Parasitic diseases still pose major obstacles to healthy growth and socio-economic development in developing countries. Some of the diseases, such as malaria, are life threatening and are the leading cause of mortality in endemic countries. Others, particularly those classified as neglected tropical diseases (NTDs), cause debilitating symptoms. The NTDs include parasitic diseases such as onchocerciasis, schistosomiasis, lymphatic filariasis and soil transmitted helminthiasis (STH), as well as bacterial diseases; leprosy and trachoma. NTDs are chronic and hinder healthy growth in children and also

significantly reduce the productive life of adults. The effects of these diseases are further magnified in the context of the subsistence economies of rural communities where they are believed, by many, to be the major cause of the poverty and disruption in social stability and economic progress. NTDs have been identified as ‘targets of opportunity’ in the effort to improve global health, while creating more vigorous economies and a better quality of life in some of the world’s poorest countries.

During the internship period the undertaken research methods related tasks were anchored on the on-going NTDs’ research projects.

1.1 Problem statement

The availability of a large variety of both elementary and complex statistical methods and the quick pace of change and development in statistics implies that researchers' statistical knowledge may be limited for them to be independent. It is unrealistic to expect researchers to solve the statistically-related challenges that arise in the course of their investigations on their own. As a result, incorporating the input of a research methods specialist in research projects will greatly improve the quantity and quality of scientific outputs from their work.

1.2 Objectives

1.2.1 General objective

To enhance the quality of scientific outputs from research in NTDs at ESACIPAC (KEMRI) by providing research methods support

1.2.2 Specific objectives

1) To improve the quality of scientific outputs through provision of research methods consultancies

2) To improve the quality of data from NTDs research projects through provision of data management support

3) To improve the quality of scientific outputs from NTDs' research projects by providing data analyses services

CHAPTER 2: LITERATURE REVIEW

2.1 Need for a Research Methods Specialist to Participate in Research/Consultancy

A research methods specialist may be consulted by researchers on a variety of challenges arising during the research process. Belle (2008) mentions that a good consulting session should begin with a problem context and definition, move to resolution and solution, and conclude with a summary and allocation of responsibilities. In addition, Belle (2008) contends that good statistical practice includes writing a brief summary of the session which, typically, includes description of the study area and problem, statistical issues, decisions and recommendations, and action items.

Batanero (2001) contends that the consultancy sessions offer the unique opportunity of teaching clients with their own data and examples and also an opportunity for consultants to change "reactive" researchers, who approach the consultant only when the data have been collected, to "proactive or collaborative" researchers, that is, clients who count on the statistician or a research methodologist from the very beginning of their research.

According to Ader *et al* (2008), a consultant may be called upon to give advice on a range of issues. These may include advising on: the initial design of the study (e.g., advising and characterizing the problem in the first place), how to: collect the data so as to most accurately and efficiently answer the research questions, construct measuring instruments for data collection, analyze the data, and interpret the results. Often, at conceptualization phase of the research project, the consultant will also be called in for advice on the broader methodological question on the transition of a substantive problem into an appropriate research design. Usually the consultation extends to issues related to the power of the study and the required sample size whereby the consultant's role is to look for a sound sample size given the constraints on the study.

Several secondary issues may also form part of the consultation session. Ader *et al* (2008) states that these involve issues such as how to cope with poor quality or missing data, whether it is legitimate to conduct intermediate analyses, and on how to best present the results to various audiences. In most cases, the consultant will be involved in generating reports and thus develop some technical passages himself because the client does not know the technicalities of the methods used. In some cases, the least time consuming approach, for both the client and the consultant, is for the consultant to write the methods section and large parts of the results and make a list of the points that should be included in the discussion and conclusion (Ader *et al*, 2008).

2.2 Data management

Research projects often involve the collection of a large volume of data which have to be processed, analyzed and the findings prepared for publication. For this sequence to proceed smoothly, the project requires a well-defined system of data management which includes an overview of the flow of data from research subjects to data analysts (Schoenbach, 2000). Chege and Muraya (2009) define data management as the process of designing data collection instruments, looking after data sheets, entering data into computer files, checking for accuracy, maintaining records of the processing steps, archiving for future access. It also includes data ownership and responsibility issues. The purpose of the data management system, according to Schoenbach (2000), is to ensure:

i. High quality data which, in turn, ensures that the variability in the data derives from the phenomena under study and not from the data collection process

ii. Accurate, appropriate, and defensible analysis and interpretation of the data

Good data management practice is a hallmark of scientific integrity. As a result, high standard of data quality in research databases should be maintained at all times (Peat and Burton, 2005). Scientists spend a great deal of time preparing data for analysis; converting data to suitable formats, merging data sets in different files, and summarizing data from field measurements. The time spent in this pre-processing step can be greatly reduced if

data are properly managed (Chege and Muraya, 2009). The following is an outline of the main stages of the data management process (Statistical Services Centre (SSC), 1998):

i. The raw data have to be entered into the computer, and checked;

ii. The data have then to be organized into an appropriate form for analysis;

iii. Archive data making it available in the subsequent project phases, and afterwards.

The key steps followed in research data management are summarized in Figure 1.

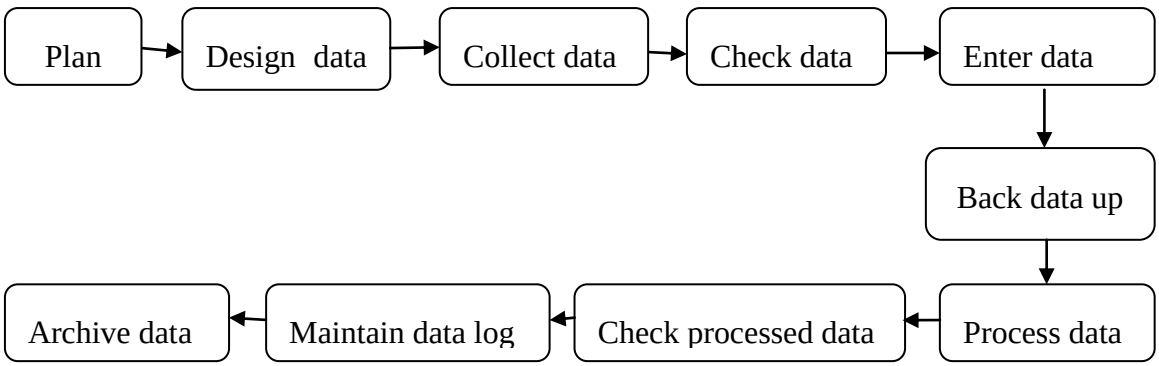


Figure 1: Data management processes (Chege and Muraya, 2009)

Before it can be analyzed, data must be collected, reviewed, coded, computerized, verified, checked, and converted to forms suited for the analyses to be conducted. The process must be adequately documented to provide the foundation for analyses and interpretation. To

preserve data integrity, the standard principle of data management, is to maintain a single "master copy" of the data with all the errors arising being corrected in the master copy (SSC, 1998). An audit trail should be maintained documenting all the changes made, who made them, and where, when, and how the changes were made. Schoenbach (2000) adds that audit trails are important for responding to or recovering from: (1) legal challenges, (2) procedural issues, (3) minor problems, and (4) disaster.

Part of data management is taking care of the security concerns which include: (1) legal, (2) safety of the information, (3) protection from external sources, (4) protection from internal sources. However, Schoenbach (2000) cautions that there can be an inverse relationship between security and accessibility/usefulness of the data. While abuse is more salient, accidental problems are more common. Typical preventive measures, recommended by Peat and Burton (2005), are removal or isolation of information that identifies research subjects (to protect confidentiality), redundancy, and backups (to protect against human and machine malfunction).

It is especially important to know the range and distribution of each variable and whether there are any outlying values or outliers so that the statistics that are generated can be explained and interpreted correctly. A considered pathway for efficient data management before beginning statistical analysis, as outlined by Peat and Burton (2005), is as follows:

i. Obtain the minimum and maximum values and the range of each variable

ii. Conduct frequency analyses for categorical variables

iii. Ascertain normality of continuous variables, e.g., using box plots and histograms

iv. Identify and deal with missing values and outliers

v. Re-code or transform variables where necessary

vi. Re-run frequency and/or distribution checks

The codes for missing values need to be accurately defined as such before statistical analyses commences, else, the missing values will be inadvertently incorporated into the analyses, thus, producing erroneous results. Although these values can be predefined as system missing, Peat and Burton (2005) argue that this is an unnecessary process because it requires familiarity with the coding scheme. They recommend indication of a missing value with a full stop rather than using the implausible value of 9 or 999. The impact of missing data is magnified for analyses involving large number of variables, since many analytical procedures require omitting any observation that lacks a value for even one of the variables in the analysis (Academic Computing Services, 2000).

Before analysis is started continous variables may be converted into ordinal categorical variables. Such data reduction activities take place even at the analysis stage (Schoenbach, 2000). Data reduction involves deciding whether and how continuous variables can be grouped into a limited number of categories and whether and how to combine individual variables into scales and indexes. It is driven by the need to derive conceptually more meaningful variables from individual data items (Academic Computing Services, 2000).

Examination for extreme values using range checks and graphical representations (boxplots, histograms, scatter plots and diagnostic plots) is a crucial preliminary step in the screening of data (Bowers, 2008). The visual impact of a graph is informative and increases the understanding of the data (Academic Computing Services, 2000). Also, a visual inspection of the data is quite informative in gaining impressions about potential outliers. Outliers may meet one or both of two possible criteria. First, outliers should be checked to the original data forms to verify accuracy of them. For an outlier to be truly an outlier it must make substantive sense. There are formal statistical tests for outliers. These tests are designed to identify those values that may unduly influence a statistical analysis. The analysis can be repeated with and without the outlier data to assess the impact of the outlier on the analysis (Feinstein, 2002). Alternatively, the analysis can be repeated using (non-parametric) statistical procedures that are not affected by outliers and the results compared to parametric procedures – or non-parametric procedures can be used completely

(Harris and Taylor, 2003). Harris and Taylor (2003) also propose that outliers may be replaced with a missing value, but then the observation is lost with regards to the analysis (and in a mathematical modeling procedure, the entire observation is unused). If the outlier is a legitimate value, then simply deleting it is a questionable procedure (Academic Computing Services, 2000). Categorical procedures, in which a variable is first categorized into groups will often be unaffected by extreme values. Also, the logarithmic transformations are particularly useful for data that have a right skew or high-value outliers while square root transformations are useful adjusting data that have a left skew or low-value outliers (Feinstein, 2002).

The last stage in data management involves archiving of the data and all other materials and resources related to the study. Concern about scientific misconduct and fraud continues to increase, and investigators have the responsibility to maintain documentation to allay any such charges should they arise. Increasingly, journals are requiring that data (and supporting documentation) be retained for several years following publication (Schoenbach, 2000).

2.3 Data analysis

Analyzing the data is the process of extracting useful information from a data set by turning the raw observations into summaries that can be interpreted (Shia, 2001). The data

can be split into two parts; pattern and residual. The aim of the data analysis is to describe the pattern which is the underlying structure or shape of the data. Residual forms the remaining, unexplained variation. There should be no pattern in the residual part of the data and if there is, this is an indication that some effect has been forgotten, perhaps due to the layout, treatments or measurements (Richardson-Kageler, 2009).

Batanero (2001) argues that computers make it easy to perform statistical analysis and often carry out complicated calculations, without understanding why they are needed or if they are needed at all, and without thinking about possible alternative methods of analysis. An ill-thought-out analysis process can produce incompatible outputs, overlook key findings and fail to pull out the subsets of the sample where clear findings are evident (SSC, 2001). Appropriate methods for analysis depend on the objectives, the study design and the nature of the observations (Richardson-Kageler, 2009).

A prerequisite to proper data analysis is the formulation of analysis objectives which are determined by, but more specific than, the overall research objectives and evolve during the research as one gains insights and experience (Richardson-Kageler, 2009). Analysis of data starts exploratory and descriptive analysis (EDA) which displays the main patterns in the data with summary tables and graphs and allow one to tentatively meet many of the analysis objectives (SSC, 2001). It can lead to additional data collection if this is seen to be needed or savings by stopping collecting data when a conclusion is already clear, or

existing results prove worthless. Richardson-Kageler (2009) highlights limitations of EDA which include:

i. Only simple patterns can be investigated, e.g., one can look at how y varies as x varies by plotting y against x. But what if there are several x's, all to be considered simultaneously?

ii. There has been no consideration of the uncertainty in any of the summaries that are used to interpret the data. Yet we know there is variation in the observations, so there is uncertainty in the results

To address these problems formal analysis is necessary. Estimating characteristics of a population of interest, from a sample is a fundamental purpose of statistical work, whether the activity is an observational or monitoring study, a survey or an experiment and formal analysis serves this purpose (SSC, 2006). Formal or confirmatory analyses add information about uncertainty and allow one to disentangle complex patterns by generating the summary findings, relationships, models, interpretations and narratives from the data (Richardson-Kageler, 2009).

Once the analysis is complete it is important to revisit the objectives to ensure they have been fulfilled. Moreover, the analysis done may not be the same as that planned in the

original project proposal thus the need to revisit the objectives and relate them to the results obtained (Richardson-Kageler, 2009).

CHAPTER THREE: METHODOLOGY

3.1 Participating in research/consultancy

Participation in research was achieved through provision of consultancy services to the scientists working in ESACIPAC. The challenges posed were documented, as presented by the scientists, after which a solution was provided on the spot or provided at an agreed later date depending on the complexity of the problem.

The consultancy sessions involved sample size determination for a cross sectional and a comparative longitudinal study as well as distribution of the resultant sample sizes to the study clusters using probability proportional to clusters' population sizes.

Other consultancies done involved carrying out statistical tests and supporting proposal development by writing sections of it that require statistical input, that is, sections on sample size determination, data management and statistical analysis.

3.2 Data management

Data management support was provided for a number of studies and reported here is data for a study, "*Adverse effects associated with mass drug administration of praziquantel and albendazole in Mwaluphamba location, Kwale*". Data collection had already been done so research support was offered from the data entry phase onwards.

The main variables in the data were: ID NO, DATE, SCHOOL, NAME OF THE PUPIL, AGE, CLASS, SEX, ADVERSE EFFECT (AE), TYPE OF AE, AE(OTHER) , DURATION OF AE, SEVERITY OF AE, AE/DRUG RELATION, ACTION TAKEN.

Double data entry was done using MS Excel which involved two data entry personnel working independently on the same data set. Supervision of the data entry clerks were done regularly. The spreadsheets for data entry were designed by the research methods intern and this involved setting up the relevant validation checks and facilitating data auditing.

Double data entry comparisons were also done using Epi Info and, using the resultant HTML report, the arising discrepancies were corrected. A MS word document containing the metadata was also prepared. The metadata included dates of data collection and data entry, title of the study and its objectives and what each variable in the spreadsheet represented. Data was finally stored in flash disks and external hard drives.

3.4 Data analysis

Analyses were always done under the tutelage of two senior statisticians (Mr. Muniu and Mr. Mwangi, Centre for Public Health Research (CPHR), KEMRI). They provided guidance on how to approach the analysis, presentation of statistical outputs and reporting of the findings.

Data were analyzed for a cross-sectional survey entitled, "*Factors contributing to re-infection with S. mansoni among primary school children: a case study of cohort schools in Mwea irrigation scheme*". The main outcome was re-infection status. To match this requirement, the original data which reflected the number of ova/cysts observed microscopically was coded appropriately. The data were then explored for cohort baseline characteristics by performing data validity checks, data completeness, and a check for outliers by a description of all the variables through simple tabulations, graphical plots (histograms, normal quantile to quantile (Q-Q) plots and detrended Q-Q plots) and simple summaries. Analyses involved univariate and multivariate analysis and modelling of re-infection with schistosomiasis using logistic regression approach.

CHAPTER 4: RESULTS AND DISCUSSIONS

4.1 Participation in research/consultancy

4.1.1 Activities/Results

4.1.1.1 Sample size calculation for a descriptive cross sectional study

Sample size determination was done for a descriptive cross sectional study, "Factors contributing to re-infection with *S. mansoni* among primary school children: a case study of cohort schools in Mwea irrigation scheme, Central Kenya (Mbinya, 2011)". The minimum sample size required for the study was computed based on one of the specific objectives, "To determine the prevalence of *S. mansoni* among selected primary school children in Mwea irrigation scheme". The following equation was used for sample size computation:

$$n = \frac{Z^2_{1-\alpha/2}\, p\,(1-p)}{d^2}$$

Equation **1 (**Lemeshow, 1990)

Where: $Z_{1-\alpha/2}$ = Standard normal deviate corresponding to 95% confidence level (1.96)

P = proportion with characteristic of interest

d = Margin of error

The desired width of the 95% confidence interval was 10%. Assuming a prevalence of 50%, the minimum required sample size required for this study was 386.

Further, the sample size for each study school and, also, for each class in those schools, were determined using probability proportional to the size (PPS) approach (Table 1).

Table 1: Sample sizes for the three schools and sample size for the classes

School	**Population/sample**	**Overall**	**Class 4**	**Class 5**	**Class 6**	**Class 7**
Mokou	All (N)	263	69	68	70	56
	Sample(n)	134	35	35	36	28
Mbui Njeru	All (N)	289	75	67	85	62
	Sample(n)	147	38	34	43	31
Mianya	All (N)	206	65	65	35	41
	Sample(n)	105	33	33	18	21
Total	All (N)	758	209	200	190	159
	Sample(n)	386	106	102	97	80

4.1.1.1 Sample size determination a comparative longitudinal study

The approach used in sample size calculation and the corresponding narrative that was done by the RM intern was as follows:

> From epidemiological studies, it is estimated that the infection rates, with soil transmitted helminthiasis (STH), among school age children, six months after de-worming, will be at least 20%. The integrated intervention will be assumed to have a significant contribution if it can lower the re-infection rate by 10 percentage points. To be able to detect whether integrated intervention results in a significantly lower re-infection rate than de-worming alone at the 5% level of significance with 90% power, a minimum sample of 266 in each study site is required.

The formula for estimating the sample size when comparing two proportions with equal sample sizes is:

$$n = \frac{\{Z_{1-\alpha/2}\sqrt{2p(1-p)} + Z_{1-\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\}^2}{(p_1 - p_2)^2}$$ Equation **2 (**Feinstein, 2002)

Where,

$Z_{1-\alpha/2}$ = Standard errors for the mean corresponding to 95% confidence level (1.96)

$Z_{1-\beta}$ = Power of the test (1.282)

P_1 = Re-infection rate in village with integrated intervention (10%)

P_2 = Re-infection rate in village with de-worming only (20%)

P = the mean of p_1 and p_2 (15%).

Since the study was based on a cluster sampling approach, the sample size was increased by an estimated design effect of 1.2 and the resultant minimum required sample size for each study site was 320. Therefore, a total of 640 participants was the minimum required enrolment for this study. However, further deliberations with the investigator revealed that the sample size also needed to cater for the participants who may drop out of the study during the follow-up period. Hence, assuming a response rate of 90%, the final minimum required sample size for the study was estimated to be 720.

4.1.1.2 Other contributions in writing the proposal

Contributions were made on drafting data management and analysis section of the proposal. This section requires statistical input. The section of the proposal was, thus, written as follows:

Data management and analysis

During data collection questionnaires will be checked regularly to rectify any discrepancies, logical errors or missing values. Quantitative data management and analysis approaches will be applied in data processing and analysis. Double data entry will be done using Microsoft Access computer package. Verification of data entry will be done through the use of Epi Info version 3.4.3. After cleaning the data, analysis will be done by the use of Statistical Package for Social Sciences (SPSS). The analysis of the data will involve computing appropriate descriptive statistics, including Chi-square, logistic regressions, odds ratios and their corresponding 95% confidence intervals. The level of statistical confidence will be $p < 0.05$ (Proposal by Mwobobia *et al,* 2011)**.**

4.1.1.3 Calculation of 95% confidence intervals

As part of the preparation of an abstract, **"**Insecticide treated nets based malaria control in Kenya: a micro-level coverage and use in a district under national malaria control programme (Mwobobia *et al,* 2011)", one of the reviewers requested additional information on the results section, that is, inclusion of 95% confidence intervals(C.I).

Out of a total of 1446 under fives, 29.3% (**95% C.I:** ……) were reported to have slept under a net … The corresponding proportion for a total of 1719 women of child-bearing age was 24.1% (**95% C.I:** ……).

Using equation 3, the 95% confidence intervals were determined for the under fives and women of child-bearing age as 25.0-33.7 and 20.0-28.3 respectively.

$$p \pm 1.96\sqrt{[p(1-p)/n]}$$ Equation **3 (**McDonald, 2009**)** where;

p = the proportion of respondents with the characteristic of interest

n= number of respondents with the characteristic of interest

4.1.1.4 Evaluating the differences in distributions of two groups of HIV positive patients

The consultation involved determining whether there is a difference in the distributions of two groups of HIV positive patients with respect to the various species of cryptosporidium, a protozoon known to cause diarrhoea. The groups were: patients with diarrhea (symptomatic) and those without (asymptomatic). The consulting scientist had presented his findings in the scientific committee meeting and questions were raised over his statistical results which had p-values of 1. He approached the intern for assistance.

The study involved 156 participants attending a clinic (HIV positive patients only) in one of the local hospitals. The patients were grouped into two: symptomatic (presenting with diarrhoea, n=70) and asymptomatic (no diarrhoea, n=86). Faecal specimens from the participants were examined for the presence or absence of the cryptosporidia ova. The positive specimens were examined further by molecular techniques and cryptosporidia species identified.

The numbers of specimens found positive were 30 and 27 for the symptomatic and symptomatic group respectively. Table 2 presents the findings from the entire study including the species of cryptosporidia identified by molecular technique.

Table 2: A summary of the results from the study

Cryptosporidia	**Species**	**Diarrhoea (n=70)**	**No diarrhoea (n=86)**
Presence	*C. hominis*	15	18
	C. parvum	7	9
	C. canis	2	1
	C. melagridis	2	0
	C. muris	1	2
Absent	None	43	56

The question was posted on the research methods course website and assistance was obtained from the students' fraternity as well as from the course facilitators. The problem was the scientist was doing the chi square test without meeting one statistical assumption for the test - using categories with values less than 5. The concerned categories were consolidated into one group named 'Other'. The other alternative was to use Fisher's exact test or better still use a logistic model. The ultimate conclusion was, "from the chi square test carried out, that the data provided no evidence that cryptosporidia species were significantly associated with the incidences of diarrhoea in HIV positive patients".

To ensure that questions arising at the post-analysis stage are addressed satisfactorily, it would be imperative to institute policies that enhance data accessibility in future even after the investigators leave the organization for one reason or the other. This includes aspects of ownership whereby data belongs to the organization a scientist is working for and/or funding the project.

Before conducting a statistical test, it is essential that one is conversant with the statistical assumptions that form the bases of such statistical tests and, more importantly, to ensure that they are met. This can be easily achieved by consulting a research methods specialist.

4.1.2 Discussion

4.1.2.1 Sample size calculation for a descriptive cross sectional study

Majority of the studies encountered at ESACIPAC, during the internship period, dealt with public health-related issues. Their objectives were mainly to describe, with means or proportions, one or more characteristics in a particular study group. The studies were usually of cross-sectional in design and by extension, almost always, involved cluster sampling. In such studies, sample size is an important consideration because it affects how precise the observed means or proportions are expected to be (Eng, 2002). Knowledge of the sample size requirements for a study plays a role in proper execution of the study and proper utilization of resources.

In the consultations, involving determination of sample size requirements, there are pertinent questions that must be addressed. Some of the questions that always arose during such consultations included:

i. Do you want to learn about a mean? A mean difference? A proportion? A proportion (risk) ratio? An odds ratio?

ii. Do you want to estimate something with a given precision or do you want to test something with a given power?

iii. What type of sample(s) will you be working with? A single group? Two or more independent groups? Matched pairs?

The sample size (n) for a cross sectional study is based on three major parameters, as shown in equation 1. These parameters are: the desired margin of error, desired confidence level from which the corresponding standard normal deviate is determined and the proportion with characteristic of interest. The investigator was interested in the proportion of children re-infected with *S. mansoni* and the corresponding 95% confidence interval for the true proportion. The desired level of precision required was such that a 95% confidence interval was to be no wider than 10% (i.e., 0.1). The standard normal deviate for the 95% confidence interval is 1.96. The margin of error was 0.05, which is half of the width of the confidence interval (0.1). Finding the proportion of re-infected children was the ultimate goal of the study and, thus, the true value remains unknown until the study is finished. The major challenge was in deciding the proportion to use in calculating n. Israel (2009) recommends use of an assumed proportion (p) of 0.5 in such instances as this is the value that gives the maximum variability. In other words, the "worst case scenario" occurs when p=0.5 and it leads to the largest value of n. A value of 0.5 was adopted for the computation of n.

Chadha (2006) recommends that part of the conversation in sample-size planning should centre on the consequences of getting it wrong: What if we find ourselves wanting a

follow-up study? How much will that set us back? Can we budget for it? Answers to such questions will help to decide how liberal or conservative we need to be in sample-size calculations. This was a vital consideration in the calculation of sample sizes for the cross-sectional study involving Mwea schools whereby the investigator had to increase the sample size by approximately 10%, which is from 386 to 420. This was because the study involved recruiting the study participants in the first day of the field work and the study participants were requested to bring faecal specimens the following day. Only those who presented the faecal specimens were interviewed. Thus, the surplus catered for the children who could not bring the specimens for one reason or the other. The same consideration also proved essential in the comparative longitudinal study sample size.

Sample size is but one aspect of study design and a lot of questions must be asked and answered. Lenth (2001) provides such examples and raises such questions as: Exactly what are the goals? What is the response variable, how do you plan to take measurements, and are there alternative instruments? What is your estimate of the non-response rate? What are the important sources of variation? How can we design the study to estimate efficiently? What is the time frame? What are the other practical constraints? Most these questions can only be answered by the concerned researcher as he is well versed with the subject matter. It requires care in eliciting scientific objectives and in obtaining suitable quantitative

information prior to the study. Thus, successful resolution of the sample-size problem requires the close and honest collaboration of statisticians and subject-matter experts.

Both Equation 1 and 2 are based on two statistical assumptions according to (Rosner, 2000, Fleiss, 1981). First, it is assumed that the selection of individuals is random and unbiased. Second, the decision to include an individual in the study cannot depend on whether or not that individual has the characteristic or outcome being studied. Both of these assumptions are required not only by the sample size calculation method, but also by the statistical (parametric) tests that are performed later. Different methods for determining sample size are required for nonparametric statistics such as the Wilcoxon rank sum test (Frison and Pocock, 1992).

In these consultations, the main role of the research methods intern was eliciting the information on the matter under study from the investigator which helped in determining some of the parameters required in calculation of sample size and also determining if the statistical assumptions are fulfilled. It was observed that skills in thinking carefully about sources of variation, and in estimating them, are other important reasons why a research methods specialist should be involved in sample-size planning.

4.1.2.2 Contributions in the development of a proposal for a comparative study

The longitudinal comparative study involved estimating the proportions of the population with the characteristic of interest followed by comparisons between the two populations. Equation 2 represents the formula for estimating the sample size when comparing two proportions with equal sample sizes. As indicated in the equation, an appropriate sample size in a comparative study generally depends on five study design parameters: minimum expected difference (effect size), estimated measurement variability, desired statistical power, significance criterion, and whether a one- or two-tailed test is to be used.

The minimum expected difference, also called effect size, is the smallest measured difference between comparison groups that the investigator would like the study to detect (Eng, 2002). The setting of this parameter is subjective and is based on experience with the problem being investigated, results from pilot studies and, also, a literature review can also guide the selection of a reasonable minimum difference. In this study, it was based on a combination of experience of the investigator on the subject of interest, literature from epidemiological studies with some facilitation from the research methods intern. Lenth (2001) contends that there exists a trade-off between a feasible sample size and adequate effect size. Consequently, bearing in mind that a compromise had to be reached between the desired effect size and a manageable sample size, the investigator proposed a range of effect sizes and with the help of the intern, the corresponding sample sizes were computed.

The effect size is inversely related to the sample size and, thus, as the minimum expected difference was made smaller, the sample size needed to detect statistical significance increased and vice versa. Adjustments were made on the proposed ranges of effect size with respect to the sample sizes and, eventually, the minimum expected difference was set at 10 percentage points.

Because of the variations resulting from sampling error, one cannot always be certain of obtaining a significant result of a study, even if there is a real difference. It is necessary to consider the probability of obtaining a statistically significant result. This probability is the power of the study. Chadha (2006) defines power as the ability of a study to enable detection of a statistically significant difference when there truly is one. To detect a minimum difference of ten percentage points, between the two study areas, a power of 90 percent was agreed to be satisfactory for this study. This implied that if the study were to be conducted repeatedly, a statistically significant result would be obtained nine times out of ten if the true difference was really ten percentage points or more. Lenth (2001) argues that it is as wasteful and inappropriate to conduct a study with inadequate power as it is to obtain a diagnostic test of insufficient sensitivity to rule out a disease. Nonetheless, while high power is always desirable, there is an obvious trade-off with the number of individuals that can feasibly be studied, given the usually fixed amount of time and

resources available to conduct a study (Eng, 2002). Conventionally, power of at least 80 percent is used (Ospina and Ortiz, 2001).

The significance criterion is the maximum *p* value for which a difference is to be considered statistically significant. As the significance criterion is decreased (made more strict), the sample size needed to detect the minimum difference increases. The significance criterion is, customarily, set to 0.05 (Eng, 2002) and the same was adapted for this study. `It was clear that any difference between comparison groups was possible in only one direction, that is the intervention was expected to cause a reduction in the prevalence of the schistosomiasis and STH. Analysis would, thus, involve one-tailed statistical tests. Calculating sample size based on this reduces the sample size required for detection of the minimum difference compared to two-tailed tests. Indeed, the sample size of a one-tailed design with a given significance criterion—for example, α—is equal to the sample size of a two-tailed design with a significance criterion of 2α, all other parameters being equal (Eng, 2002). This issue was not considered at the time but it was communicated later to the investigator and he incorporated the changes into his study, that is, the minimum required sample size was reduced from 720 to 320.

There are freely available softwares that can also be used to calculate sample sizes one of them being Epi Info®. When using the softwares, it is important to bear in mind some unique features as well as the limitations that may arise from using such, for instance,

Rahbar (2005) points out that Epi Info adds a continuity correction to the estimates thus modestly elevating the calculated sample sizes. In addition, it requires that you specify what the minimum odds ratio, OR = (p1/1-p1), or relative risk you want to detect.

The sample size calculations should always be considered estimates of an absolute minimum (Browner *et al*, 2001). The words ‘minimum required’ sample size and the prudence of planning to include more than the minimum number of individuals in a study, if and when possible, were aspects that were always impressed upon the consulting investigators.

There are many secondary issues that may also form part of the consultation sessions. The consultant, according to Ader *et al* (2008), will be involved in generating reports whereby he will write some technical passages himself because the client does not know the technical ins and outs. This consultation case ended up with a request to assist in drafting the relevant sections of the proposal by bringing together all the technical aspects discussed. Batanero (2001) contends that the consultancy sessions offer the unique opportunity for consultants to change "reactive" researchers to "proactive or collaborative" researchers and this proved true since the investigators decided to include the intern as a co-investigator in this study.

4.1.2.3 Design effect and probability proportional to size approach

In all the consultations concerned with sample size determination reported in this document, incorporation of the design effect on sample sizes was done since the studies involved cluster sampling design. Cluster sampling is commonly used, rather than simple random sampling, mainly as a means of saving resources. Respondents in the same cluster are likely to be somewhat similar to one another and this implies that in a clustered sample, selecting an additional member from the same cluster adds less new information than would a completely independent selection (Alecxih, Corea and Marker, 1998). Therefore, the sample is not as varied as it would be in a random sample, so that the effective sample size is reduced (Levy and Lemeshow, 1999). The loss of effectiveness by the use of cluster sampling, instead of simple random sampling, is the design effect. Design effect is the ratio of the actual variance, under the sampling method actually used, to that of simple random (US Census Bureau, 2007).

A practical problem encountered with this approach was deciding the appropriate estimates to use for the design effect. Firstly, a study will involve many variables and each variable may have a different design effect. Secondly, to the best of my knowledge, there are no published design effects for the various regions in Kenya, or estimates of intraclass correlation coefficient from which the design effects could be calculated. Intraclass

correlation coefficient is a measure of the homogeneity of elements within clusters (Rowe *et al*, 2001).

The studies also needed determination the number of elements to be picked from each cluster (schools, classes in schools and villages). The clusters were of varying sizes (populations) and thus were sampled with probability proportional to their population sizes approach. This ensured that each element in the study had an equal probability of being included in the study. Probability proportional to size (PPS) is a sampling technique in which the probability of selecting a sampling unit, e.g., village, is proportional to the size of its population. It is most useful when the sampling units vary considerably in size because it assures that those in larger sites have the same probability of getting into the sample as those in smaller sites, and vice versa (McGinn, 2004). This method also facilitates planning for field work as a pre-determined number of respondents are interviewed in each unit selected, and staff and resources can be allocated accordingly.

4.1.2.4 Evaluating the differences in distributions of two groups of patients

The difference in the two groups of patients was determined using a chi square test and the groups were found not to be different from each other with respect to the species of cryptosporidia. The challenge with the question was to reorganize the data presented by the scientist intuitively after which it was easy to realize a $\chi 2$ test would provide a solution.

Responding to this challenge which was posted on the course discussion forum, Coe (2011) contended that a chi square test (or t-test or many others) may sometimes be a useful tool in data analysis, but is very rarely a useful way of describing and framing an analysis. Typically, more progress is made by developing, fitting and interpreting a statistical model. Unfortunately, the complete data to facilitate creation of a statistical model (logistic regression model) were inaccessible since the principal investigator, who held the comprehensive data set had left the institute. This brought about data management issues, out of which the following recommendations were suggested:

i. Raw data is needed to do insightful analyses

ii. It must be archived so that it is accessible to researchers after the originators move on. It is also imperative that institutions should put in place a data management policy that ensures that data are owned by the institution and not individual scientists. This will greatly enhance accessibility and availability of data to other interested users.

iii. It must be accompanied by a detailed protocol that explains exactly how and why it was collected (Coe, 2011)

The consultancy sessions offer the unique opportunity of teaching clients with their own data and examples and also an opportunity for consultants to change "reactive" researchers

to "proactive or collaborative" researchers (Batanero, 2001). The client was able to learn from this consultation the importance of good data management practices and appreciated the need to put appropriate guidelines or even policies on data ownership and accessibility in research projects.

4.1.2.5 Calculation of 95% confidence intervals (CI) for a proportion

This was an easy task once the understanding on how to calculate standard error (SE) for a given proportion was achieved, i.e., SE = pq/n, where p= is the proportion of interest while q= 1-p and n is the sample size. Once the SE was computed it was used to construct the confidence interval based on the relationship between the two variable which according to Bowers (2003), the 95% confidence interval for the population proportion is equal to the sample proportion plus or minus $Z_{1-\alpha/2}$*standard errors (s.e) where $Z_{1-\alpha/2}$ is the standard normal deviate for the 95% confidence interval and is equal to 1.96. This implies that, s.e = $\sqrt{[p(1-p)/n}$ and equation 3 ($p \pm 1.96\sqrt{[p(1-p)/n]}$) can thus be rewritten as; $p \pm 1.96$ s.e.

Chadha (2006) defines the CI as a range of plausible values for the true value of the outcome measure. The observed value of the outcome measure gives the best estimate of the true value. Hence, to give some indication of the precision of this estimate, CI is attached to it. The confidence interval is inversely proportional to the precision of the estimate; thus, the narrower the confidence interval the greater the precision of the

estimate, and vice versa. The size of a confidence interval is related to the sample size of the study. Larger studies usually have a narrower confidence interval (Harris and Taylor, 2003). The equation for calculating the confidence limits of a proportion is based on the statistical assumption that the sample proportions are normally distributed (McDonald, 2009). Thus, the intern had also had also another critical role of checking if statistical assumptions were met apart before embarking on the calculations.

4.2 Data management

4.2.1 Activities/Results

Data management was done for the study, "*Adverse effects associated with mass drug administration of praziquantel and albendazole in Mwaluphamba location, Kwale*". The data set had a total of 7615 cases and 32 variables. The 32 variables were constituted by fourteen main variables, six of which were of the multiple response type. The main variables included: Identification number (ID NO), date (DATE), Schools (SCHOOL), Sub-location (SUB), name (NAME), age(AGE), class(CLASS), sex (SEX), Adverse effects(AE), number of days AE lasted (DAYS),severity of AE (SEVRTY), whether the AE was related to the study drugs (RELATD), action taken to mitigate the side effect

experienced (ACTON). The last six variables had multiple responses. An identification number was added to each multi-response label to differentiate them from each other, e.g., AE1, AE2, and AE3 to represent first, second and third AE experienced by the respondent.

Data entry was done in Microsoft Excel. The spreadsheet was adequate for the requirements and the structure of this data set, that is, the data had a simple structure (no hierarchies and separate tables) and data collection was at a single level thus there were no relationships to be defined. The data set was also not large considering that one sheet was enough and did not involve numerous multiple users as this would have warranted complex security and sharing needs and hence the need to utilize a database. The data entry spreadsheets were set up using Excel facilities that provide measures for minimizing data entry errors while maximizing processing efficiency. These measures constituted the following:

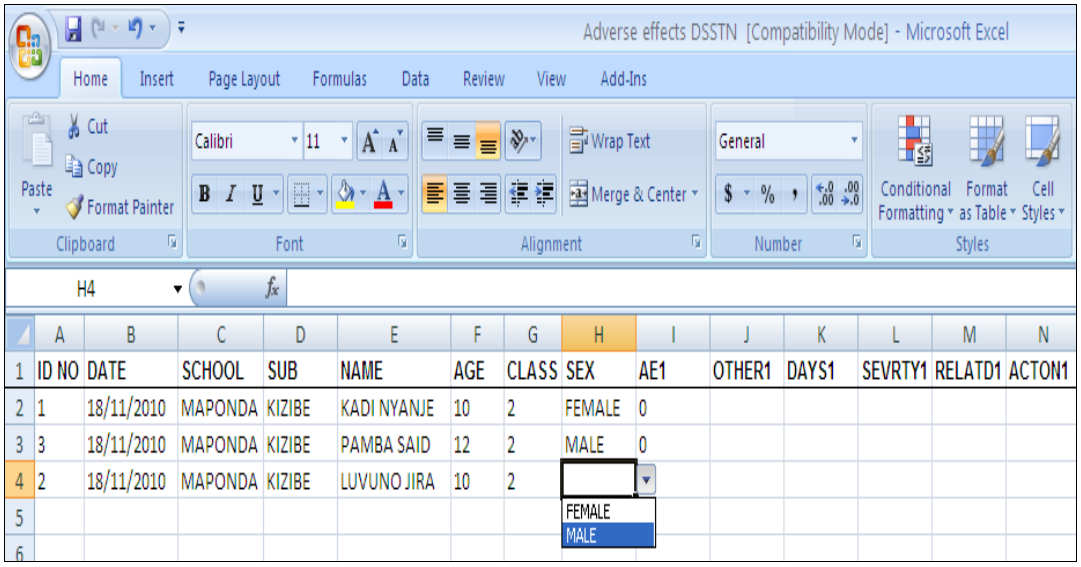
i. Freezing top row (Title row)

ii. Creating of drop down lists for the variables, e.g., for the variable SCHOOL

iii. Validation checks, e.g., for the variable CLASS

iv. Data auditing, e.g., for the variable AGE

Freezing top row kept the headings of the columns visible as one scrolled down the screen during data entry process (figure 2). To avoid typing a sequence more than once, Excel's facility for filling series and for creating drop down lists were utilized. The SCHOOL and SUB columns had a repeating list of text strings. Each text string was typed just once and then the spreadsheet's **Fill** option (**Home → Editing** menu) used to fill the remaining cells which shared the same text string. Drop down lists were also created for variables whose text strings did not follow a definite pattern, e.g., SEX as shown in figure 2.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID NO	DATE	SCHOOL	SUB	NAME	AGE	CLASS	SEX	AE1	OTHER1	DAYS1	SEVRTY1	RELATD1	ACTON1
2	1	18/11/2010	MAPONDA	KIZIBE	KADI NYANJE	10	2	FEMALE	0					
3	3	18/11/2010	MAPONDA	KIZIBE	PAMBA SAID	12	2	MALE	0					
4	2	18/11/2010	MAPONDA	KIZIBE	LUVUNO JIRA	10	2							
5														
6														

Figure 2: Drop down list for SEX and a frozen top row

Validation checks were set up for variables with numerical data such as CLASS and AGE based on the information provided in the study protocols, for instance, the inclusion and exclusion criteria for the variable CLASS. The classes ranged from 1 to 7 while the age ranged from 6 to 18 years. The adverse effects experienced by the respondents were coded from 1 to 9 and a validation checks were also set up for such coded categorical variables. Figure 3 illustrates setting up of validation checks for AE1 (adverse effect 1).

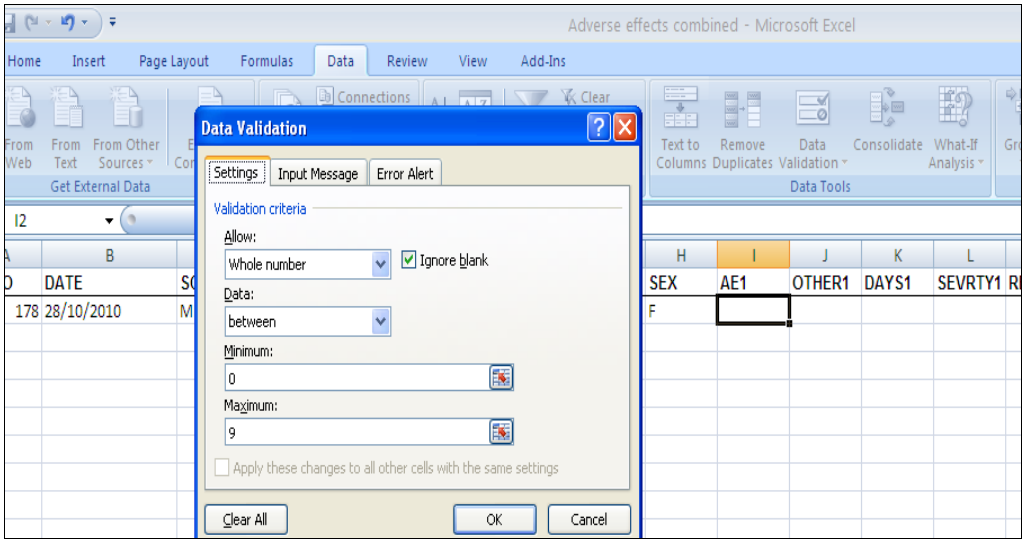


Figure 3: Setting up validation checks for AE1

The categorical variables, AE, SEVRTY, RELATD, had already been coded in the data collection forms and were entered as such in the spreadsheets. For example, severity of the adverse SEVRTY was coded as follows in the data collection form: Mild=1, Moderate=2,

Severe=3, Serious=4. Other categorical variables in the data collection form were not coded e.g., SCHOOL, SUB and SEX. They were entered into the spreadsheets as the appeared in the spreadsheets. Missing data was entered as the implausible value, 999.

A Microsoft word sheet containing metadata was maintained in the study's folder. The metadata consisted of, among other things:

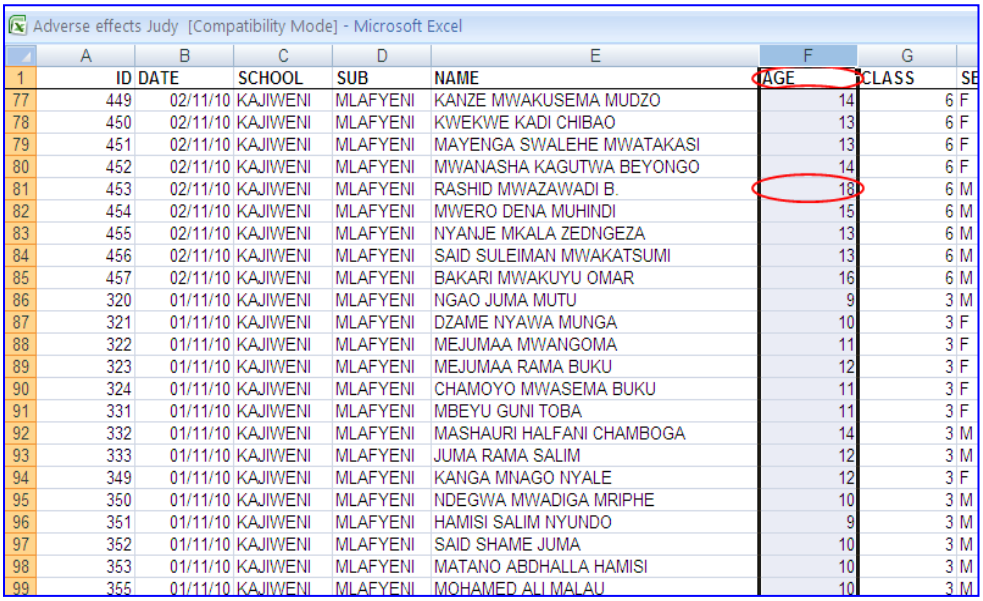
i. Study details, e.g., title, objectives, dates of data collection and data entry, etc

ii. numerical codes and their corresponding interpretations

iii. The full variable name for the abbreviated version of the variable name given in the spreadsheet with the data set, e.g. "AE" = adverse effect, "SEVRTY" = severity of the adverse effect, "RELATD" = was the adverse effect experienced related to the study drugs?

To achieve double data entry, two data entry personnel were recruited and trained after which they set out to do data entry independently. Supervision was done during the course of data entry with the supervisor visiting each of the data entry personnel at least twice in a day to check on the progress. Also reviews were done for the data entered at the end of the day by picking a number of questionnaires at random and checking if the entries made corresponded to those in the raw data form.

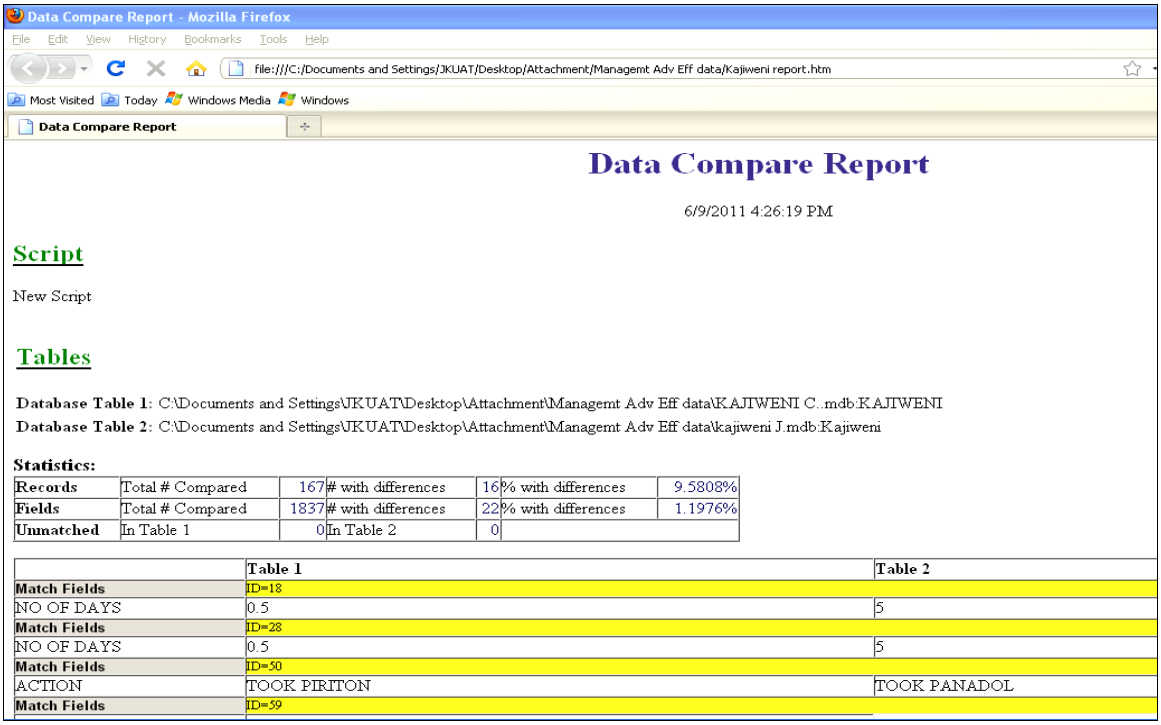
Once the data entry was complete, the data were subjected to visual inspection and any anomalies spotted were collected, e.g., blank entries and suspicious entries were counterchecked against the data collection forms and corrections done where applicable. In addition, data auditing was done for those variables whose validation rules changed during the course of the data entry process. This involved putting up new validation checks for the affected variables and running the checks and any data outside the defined range was highlighted as illustrated in figure 3.


Adverse effects Judy [Compatibility Mode] - Microsoft Excel

	A	B	C	D	E	F	G	
1	ID	DATE	SCHOOL	SUB	NAME	AGE	CLASS	SE
77	449	02/11/10	KAJIWENI	MLAFYENI	KANZE MWAKUSEMA MUDZO	14	6	F
78	450	02/11/10	KAJIWENI	MLAFYENI	KWEKWE KADI CHIBAO	13	6	F
79	451	02/11/10	KAJIWENI	MLAFYENI	MAYENGA SWALEHE MWATAKASI	13	6	F
80	452	02/11/10	KAJIWENI	MLAFYENI	MWANASHA KAGUTWA BEYONGO	14	6	F
81	453	02/11/10	KAJIWENI	MLAFYENI	RASHID MWAZAWADI B.	18	6	M
82	454	02/11/10	KAJIWENI	MLAFYENI	MWERO DENA MUHINDI	15	6	M
83	455	02/11/10	KAJIWENI	MLAFYENI	NYANJE MKALA ZEDNGEZA	13	6	M
84	456	02/11/10	KAJIWENI	MLAFYENI	SAID SULEIMAN MWAKATSUMI	13	6	M
85	457	02/11/10	KAJIWENI	MLAFYENI	BAKARI MWAKUYU OMAR	16	6	M
86	320	01/11/10	KAJIWENI	MLAFYENI	NGAO JUMA MUTU	9	3	M
87	321	01/11/10	KAJIWENI	MLAFYENI	DZAME NYAWA MUNGA	10	3	F
88	322	01/11/10	KAJIWENI	MLAFYENI	MEJUMAA MWANGOMA	11	3	F
89	323	01/11/10	KAJIWENI	MLAFYENI	MEJUMAA RAMA BUKU	12	3	F
90	324	01/11/10	KAJIWENI	MLAFYENI	CHAMOYO MWASEMA BUKU	11	3	F
91	331	01/11/10	KAJIWENI	MLAFYENI	MBEYU GUNI TOBA	11	3	F
92	332	01/11/10	KAJIWENI	MLAFYENI	MASHAURI HALFANI CHAMBOGA	14	3	M
93	333	01/11/10	KAJIWENI	MLAFYENI	JUMA RAMA SALIM	12	3	M
94	349	01/11/10	KAJIWENI	MLAFYENI	KANGA MNAGO NYALE	12	3	F
95	350	01/11/10	KAJIWENI	MLAFYENI	NDEGWA MWADIGA MRIPHE	10	3	M
96	351	01/11/10	KAJIWENI	MLAFYENI	HAMISI SALIM NYUNDO	9	3	M
97	352	01/11/10	KAJIWENI	MLAFYENI	SAID SHAME JUMA	10	3	M
98	353	01/11/10	KAJIWENI	MLAFYENI	MATANO ABDHALLA HAMISI	10	3	M
99	355	01/11/10	KAJIWENI	MLAFYENI	MOHAMED ALI MALAU	10	3	M

Figure 4: Data auditing for AGE

Double-data entry comparisons were done using Epi Info®. This is a public domain software and is able to generate reports based on the global ID variable which makes it easy to go back to the raw data forms and verify the conflicting entries. Epi Info® data comparison utility recognizes database files (.db format) only. The spreadsheets were thus exported to Microsoft Access and converted into databases whereby they were saved in .db format. Comparisons were made and the resultant discrepancies produced as HTML reports, as shown in figure 5. Based on the reports, corrections were made and the comparisons repeated until no difference was detected between the two sets of data.


Data Compare Report - Mozilla Firefox

File Edit View History Bookmarks Tools Help

file:///C:/Documents and Settings/JKUAT/Desktop/Attachment/Managemt Adv Eff data/Kajiweni report.htm

Most Visited Today Windows Media Windows

Data Compare Report

Data Compare Report

6/9/2011 4:26:19 PM

Script

New Script

Tables

Database Table 1: C:\Documents and Settings\JKUAT\Desktop\Attachment\Managemt Adv Eff data\KAJIWENI C..mdb:KAJIWENI
Database Table 2: C:\Documents and Settings\JKUAT\Desktop\Attachment\Managemt Adv Eff data\kajiweni J.mdb:Kajiweni

Statistics:

Records	Total # Compared	167	# with differences	16	% with differences	9.5808%
Fields	Total # Compared	1837	# with differences	22	% with differences	1.1976%
Unmatched	In Table 1	0	In Table 2	0		

	Table 1	**Table 2**
Match Fields	ID=18	
NO OF DAYS	0.5	5
Match Fields	ID=28	
NO OF DAYS	0.5	5
Match Fields	ID=50	
ACTION	TOOK PIRITON	TOOK PANADOL
Match Fields	ID=59	

Figure 5: Double data entry comparison report produced by Epi Info®

Further validations were done by generating an alternative layout for the data in form of summaries using the pivot table facility in Excel, e.g., frequency tables and contingency tables for SCHOOL and SEX. The data was then exported to the statistical package (SPSS version 11.0). In preparation for analysis, coding was done for the categorical variables which had not been coded in the data collection forms. For instance, the variable RELATD was coded as: Yes=1 and No=2, while SEX was coded as MALE=1 and FEMALE=2. For the coded responses, the numerical codes were given appropriate descriptions were added e.g. for the variable SEVRTY the numerical codes (1, 2, 3 and 4) were given the corresponding descriptions as specified in the data collection form, i.e., Mild=1, Moderate=2, Severe=3 and Serious=4. Also grouping responses was done in an intuitive way so that are more meaningful to the study, for example, the continous variable AGE was converted into an ordered categorical variable by grouping the ages and coding them, that is: 1 = 6-10 years, 2 = 11-15 years, and 3 = > 15 years. The missing value code (999) was defined as system missing in SPSS to ensure that it was not incorporated in the statistical analysis.

The data was saved in flash disks, CDs and in an external hard disk as a backup.

4.1.3 Discussion

Chege and Muraya (2009) define data management as the process of designing data collection instruments, looking after data sheets, entering data into computer files, checking for accuracy, maintaining records of the processing steps, and archiving it for future access and also includes data ownership and responsibility issues. In this case data management started from the data entry phase because the data had been collected prior to the arrival of the intern. The Excel's features were employed to facilitate simple and reliable data entry. The rationale here is that the simpler the data-entry process, the more reliable will be the data that are entered. To keep the data entry process simple, the process was designed in a way that ensured that the data entry personnel entered the data in the spreadsheets as it appeared in the data collection forms, e.g., if the data was not coded, for example SEX, it was entered as such and coding done later.

Before data entry was started, the spreadsheets were set up with appropriate measures for minimizing data entry errors and maximizing processing efficiency. This preparatory phase took time but the extra effort was justified by the resultant high quality data. High standard of data quality in research databases should be maintained at all times because good data management practice is a hallmark of scientific integrity (Peat and Burton, 2005). Such efforts, according to Schoenbach (2000), ensure that the variability in the data derives from

the phenomena under study and not from the data collection process. Moreover, it also facilitates accurate, appropriate, and defensible analysis and interpretation of the data.

The columns with validation checks did not allow leaving of blank cells during data entry. In cases of missing entries in such variables, the code for missing values for those variables in the data set was defined as 999. Unfortunately, this implausible value fell out of the range checks set up as part of the data validation schemes. These resulted in changing, or even complete abolishment, of the validation checks. This necessitated data auditing to be done once data entry was completed. The codes for missing values needed to be accurately defined as such before statistical analyses commences, else, the missing values will be inadvertently incorporated into the analyses, thus producing erroneous results. The value, 999, was thus predefined as system missing in SPSS before embarking on statistical analysis. Peat and Burton (2005) argue that this is an unnecessary process because it requires familiarity with the coding scheme and, thus, recommends indication of a missing value with a full stop rather than using the implausible value of 9 or 999. However, it was realized it does not work with Excel particularly if the validation checks are in place. It is only applicable if data entry is being made directly into SPSS. The impact of missing data is magnified for analyses involving large number of variables, since many analytical procedures, e.g., modelling, require omitting any observation that lacks a value for even one of the variables in the analysis (Academic Computing Services, 2000).Thus,

as part of data management it is imperative to put in place mechanisms for minimizing the number of missing values, e.g., by training data collection personnel and supervision during data collection.

It is especially important to know the range and distribution of each variable and whether there are any outlying values or outliers so that the statistics that are generated can be explained and interpreted correctly. A considered pathway for efficient data management before beginning statistical analysis as outlined by Peat & Burton (2005) is as follows;

i. Obtain the minimum and maximum values and the range of each variable

ii. Conduct frequency analyses for categorical variables

iii. Use box plots, histograms and other tests to ascertain normality of continuous variables

iv. Identify and deal with missing values and outliers

v. Re-code or transform variables where necessary

The first two aspects were addressed by the generation of pivot tables in Excel while the rest were addressed using SPSS. A visual inspection of the data is actually quite informative in gaining impressions about anomalies in the data including spotting potential outliers. For an outlier to be truly an outlier it must not make substantive sense (Bowers, 2008). Any suspected outlier was, therefore, subjected to formal statistical tests before it

was qualified as such. This process was meant to identify those values that may unduly influence a statistical analysis. The analysis can be repeated with and without the outlier data to assess the impact of the outlier on the analysis. Or, the analysis can be repeated using (nonparametric) statistical procedures that are not affected by outliers and the results compared to parametric procedures – or nonparametric procedures can be used completely (Harris and Taylor, 2003). The involvement of statistical test for outliers was a revelation of how closely linked data management is to statistical analysis. No outliers were detected in this data set but there was one suspected outlier in the variable AGE (one 'child' was aged 18). First, this potential outlier was checked against the original data forms to verify its accuracy. Harris and Taylor (2003) suggest that outliers may be replaced with a missing value, but then the observation is lost with regards to the analysis (and in a mathematical modeling procedure, the entire observation is unused). On the other hand, Academic Computing Services (2000) puts into question such an approach by contending that if the outlier is a legitimate value, then simply deleting it is a questionable procedure. Categorical procedures, in which a variable is first categorized into groups, like coding of AGE in this dataset, will often be unaffected by extreme values and may be a good solution.

AGE was categorized and coded into three ordered categories; 1 = 6-10 years, 2 = 11-15 years, and 3 = > 15 years. Such data reduction activities took place even at the analysis

stage and involves deciding whether and how continuous variables can be grouped into a limited number of categories and whether and how to combine individual variables into scales and indexes and is driven by the need to derive conceptually more meaningful variables from individual data items (Belle, 2008, Academic Computing Services, 2000).

The visual impact of a graph is informative and will increase the understanding of the data and limit the surprises that may occur (Academic Computing Services, 2000). Graphical representations of data using boxplots, histograms, scatter plots and diagnostic plots were found to be extremely useful in the examination of the data for anomalies. Scatter plots were generated using Excel. The rest (boxplots, histograms and diagnostic plots) had to be generated using in SPSS since Excel lacks the facilities to produce them.

The last stage in data management involved archiving of the data and all other materials and resources related to the study and this was done using CDs and external hard drives. Schoenbach (2000) contends that concern about scientific misconduct and fraud continues to increase, and investigators have the responsibility to maintain documentation to allay any such charges should they arise. In addition, increasingly, journals are requiring that data (and supporting documentation) be retained for several years following publication.

4.3 Data Analysis

4.3.1 Activities/Results

4.3.1.1 Descriptive / Exploratory analysis

Data analysis was done for a study, 'Factors contributing to re-infection with *S. mansoni* among primary school children: a case study of cohort schools in Mwea irrigation scheme'. The data had 74 variables and 389 cases. Analysis was done using SPSS as specified by the investigator. SPSS (Statistical Package for the Social Sciences), as the name suggests, is a powerful statistical packages tailored to the needs of social scientists.

The dependent variable in this study was the presence or absence of *S. mansoni* in the primary school children who had been dewormed earlier. However, the dependent variable was collected as the number of ova/cysts observed microscopically and this had to be adjusted in a bid to ensure that they conformed to the objectives of the study. The independent variables included the selected demographic, socio-economic and environmental factors.

Exploratory/descriptive analysis formed the first step. This was done after checking the distributions of the variables, for instance, distribution of continous variables such as, age, cost and distance to the nearest health facility, were checked as illustrated using the

tabulated output from the variable AGE (table 3). The output also included boxplots, histograms, normal Q-Q plots and detrended normal Q-Q plots. The output was suggestive of a distribution that is different from normal as indicated by the descriptives statistics (mean and median were different). Skewness and kurtosis also had values that were different from zero (ideal value for normal distribution) as shown in table 3.

Table 3: Selected descriptive statistics for the variable AGE

Descriptives for variable AGE	**Statistic**	**Standard error**
Mean	11.69	0.83
Median	12.00	
Standard deviation	1.640	
Interquartile range	3.00	
Skewness	0.317	0.124
Kurtosis	-0.083	0.257

The histogram, for AGE, had a tail to the right which was suggestive of the distribution for the variable age being different from normal (Figure 7).

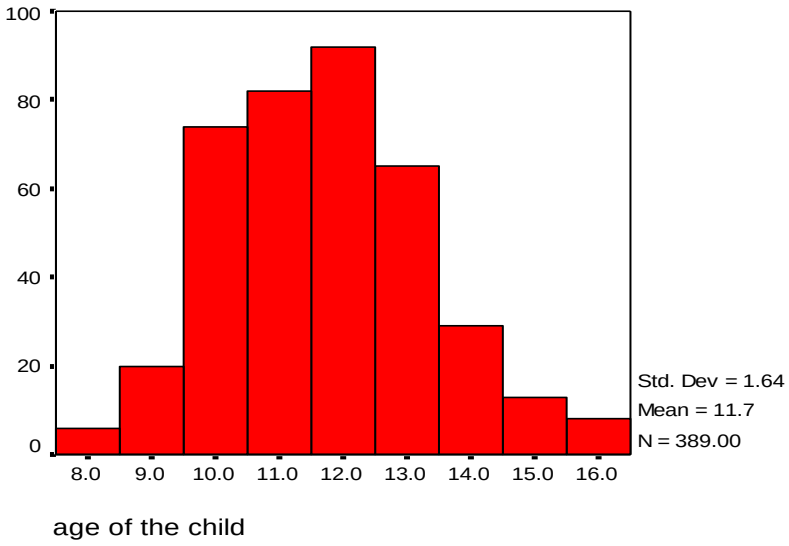


Figure 6: Histogram for the variable AGE

The distribution of the data points for the variable AGE did not fall on a straight line on the normal Q–Q plot (Figure 8). Thus, just like in the case of the histogram, it was difficult to tell whether the deviation for this variable was significantly different from normal.

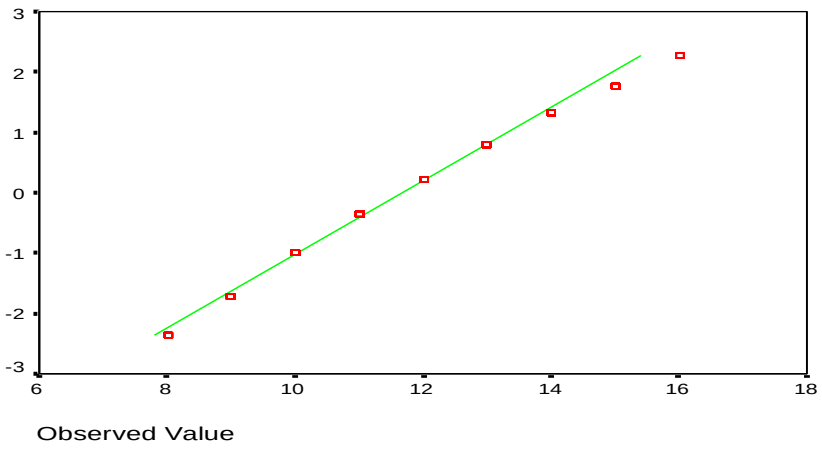


Figure 7: Normal Q-Q plot for the variable AGE

On the detrended normal Q–Q plot, the data points for the variable AGE formed a pattern that was similar to a U-shape and the horizontal line is not lying at the centre of the data which was indicative of some degree of non-normality (Figure 8).

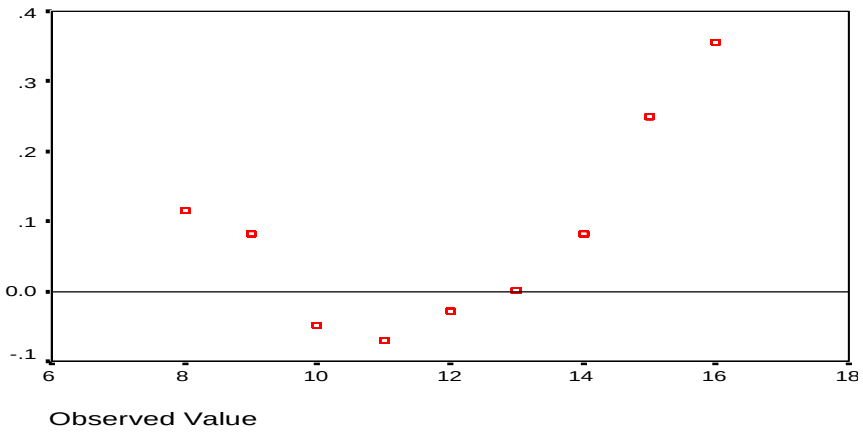


Figure 8: Detrended normal Q-Q plot for the variable AGE

The study participants were described on the basis of various characteristics; demographic, environmental and socioeconomic characteristics. Categorical variables were summarized using frequencies, rates and proportions. Continous variables which were found to be normally distributed were summarized using mean and standard deviation; else, the median and the corresponding interquartile range were used as measures of central tendency and measures of dispersion respectively. The overall prevalence of *S. mansoni* and the corresponding 95% confidence interval were computed. Prevalences were also computed by age, gender, school and location.

The economic status of the respondents' households were categorized according to their economic status ('poor/low', 'average', 'rich/high') based on a proxy obtained by scoring of the items owned by the household, i.e., ownership of items (mobile phone, television set, radio etc), presence/absence of electricity, the type of materials used to construct the house's roof and walls. The resultant economic status rankings were; poor (20%), average (60.7%) and high (18.5%).

On assessment of knowledge of the respondents, there were 13 questions regarding transmission, prevention and control as well as curability of schistosomiasis. A summary score was developed from these questions by assigning one point for each correct response and zero for each incorrect or uncertain response. A score of zero was rated as no knowledge. All the respondents who possessed some knowledge about schistosomiasis

were dichotomized, on the basis of the total knowledge score, as either Low (1- 6) or high (7 -13).

4.3.1.2 Bivariate analysis

Bivariate analyses were used to identify variables predictive of re-infection with *S. mansoni*. Crude associations between the dependent variable (presence =1, absence = 2) and each of the independent/predictor variables were explored. The statistical significance was set at $P < 0.05$. Odds ratio and corresponding 95% confidence intervals were computed for those associations that were found to be statistically significant.

Seven factors met the set criterion and they included; School, Participation in rice planting, Class, Economic activity of the household head, Main source of water for the household, No. of people per household and Economic status of the household.

4.3.1.3 Multivariate logistic regression

All factors identified to be significantly associated with re-infection by bivariate analysis were considered for multivariate analysis. The dependent variable was coded appropriately for this analysis, i.e. the dichotomous re-infection status outcome (0 = not diseased and 1 = diseased). The exit criterion was set at 0.10. Five successive iterations were done using backward conditional method retaining two factors, out of the seven; i.e., Schools and Economic status of the household. For those two variables, the adjusted odds ratio, 95%

confidence intervals and p-values were noted. The final logistic regression model was assessed for goodness-of-fit using Hosmer-Lemeshow test and it showed a reasonable fit (P= 0.270) as presented in table 4.

Table 4: Results from Hosmer- Lemeshow test

Step	**Chi-square**	**df***	**p-value**
1	4.0862	8	0.772
2	2.042	8	0.980
3	6.434	8	0.588
4	8.901	8	0.351
5	8.761	7	0.270

*Degrees of freedom

Once the analysis was completed, reporting of the analysis results was done by the investigator with the assistance of the research methods intern.

4.2.3 Discussion

Prior to beginning statistical analysis, it is important to have a thorough working knowledge of the nature, ranges and distributions of each variable (Peat and Burton, 2005). This was found to save time in the end by avoiding repetition of analyses for various

reasons. The information about the distribution of the variables helped in determining whether parametric or non-parametric tests needed to be used thus ensuring that the results of the statistical analyses can be accurately explained and interpreted. Before beginning statistical analyses of a continuous variable it is essential to examine the distribution of each of the variables for skewness (tails), kurtosis (peaked or flat distribution), spread (range of the values) and outliers (Peat and Burton, 2005). In this study, checking of the distributions was done using normality tests. However, the tests of normality do not provide any information about why a variable is not normally distributed and, therefore, they had to be supplemented with descriptive tabulations, which present the skewness and kurtosis values, and plots to facilitate identification of the reasons for the non-normality.

Parametric tests are used when a continuous variable is normally distributed. In general, parametric tests are preferable to non-parametric tests because a larger variety of tests are available and, as long as the sample size is not very small, they provide approximately 5% more power than rank tests to show a statistically significant difference between groups (Healy, 1993). Also, non-parametric tests can be a challenge to present in a clear and meaningful way because summary statistics such as ranks are less familiar to many people than summary statistics from parametric tests. Summary statistics from parametric tests such as means and standard deviations are always more readily understood and more easily communicated than the equivalent rank statistics from non-parametric tests.

The distribution of the variables was checked using tabulated results showing the descriptive statistics and graphical presentations such as normality plots, box plots and histograms giving an idea of how the data were distributed. According to Healy (1993), a quick informal check of normality is to examine whether the mean and the median values are close to one another. In the case of the variable AGE, the tabulated results on exploratory analysis showed that the mean and median had different values an indication that the variables were, probably, not normally distributed. Also tabulated, together, with summary statistics (mean, median, etc) were measures of distribution, such as skewness and kurtosis, which indicate how much a distribution varies from a normal distribution. A perfectly normal distribution has skewness and kurtosis values equal to zero. Skewness values that are positive indicate a tail to the right and skewness values that are negative indicate a tail to the left. Values between −1 and +1indcate an approximate bell shaped curve (Peat and Burton, 2005). The values of skewness and kurtosis being greater than one indicated a distribution that differed significantly from a normal, symmetric distribution.

The histograms show the frequency of measurements and the shape of the data and therefore provide a visual judgement of whether the distribution approximates to a bell shape. Histograms also show whether there are any gaps in the data, whether there are any outlying values and how far any outlying values are from the remainder of the data. For the variable age the histogram showed a slight deviation from the bell shape. The normal Q–Q

plot shows each data value plotted against the value that would be expected if the data came from a normal distribution. The values in the plot are the quantiles of the variable distribution plotted against the quantiles that would be expected if the distribution was normal (Frison and Pocock, 1992). If the variable was normally distributed, the points would fall directly on the straight line. Any deviation from the straight line indicates some degree of non-normality as observed for variable AGE.

The detrended normal Q–Q plots show the deviations of the points from the straight line of the normal Q–Q plot. If the distribution is normal, the points will cluster randomly around the horizontal line at zero with an equal spread of points above and below the line. If the distribution is non-normal, the points will be in a pattern such as J or an inverted U distribution and the horizontal line may not be in the centre of the data, e.g., the U-distribution pattern was observed for AGE.

The box plot shows the median as the black horizontal line inside the box and the inter-quartile range as the length of the box. The inter-quartile range indicates the 25th to 75th percentiles, that is, the range in which the central 25% to 75% of the data points lie. The whiskers are the lines extending from the top and bottom of the box. The whiskers represent the minimum and maximum values when they are within 1.5 times above or below the Interquartile range. If values are outside this range, they are plotted as outlying or extreme values. Any outlying values that are between 1.5 and 3 box lengths from the

upper or lower edge of the box are shown as open circles, and are identified with the corresponding number of the data base row. Extreme values that are more than three box lengths from the upper or lower edge of the box are shown as asterisks. Extreme and/or outlying values should be checked to see whether they are univariate outliers. If there are several extreme values at either end of the range of the data or the median is not in the centre of the box, the variable will not be normally distributed. If the median is closer to the bottom end of the box than to the top, the data are positively skewed. If the median is closer to the top end of the box, the data are negatively skewed.

The tabulated results and graphical presentation were only suggestive of the non-normality in distribution and the Kolmogorov-Smirnov statistic (K-S) tests proved useful as a confirmatory test. The null hypothesis for this test is that the data are normally distributed (Harris and Taylor, 2003). The SPSS analysis output on the normality tests presents the tabulated results for two tests: a Kolmogorov–Smirnov statistic with a Lilliefors significance correction and a Shapiro–Wilk statistic. A limitation of the Kolmogorov–Smirnov test of normality without the Lilliefors correction is that it is very conservative and is sensitive to extreme values that cause tails in the distribution. The Lilliefors significance correction renders this test a little less conservative. The Shapiro–Wilk test has more statistical power to determine a non-normal distribution than the Kolmogorov–Smirnov test (Stevens, 1996). The Shapiro–Wilk test is based on the correlation between

the data and the corresponding normal scores and will have a value of 1.0 for perfect normality. A distribution that passes these tests of normality provides extreme confidence that parametric tests can be used. However, a variable that does not pass these tests may not be so non-normally distributed that parametric tests cannot be used, especially if the sample size is large. This is not to say that the results of these tests can be ignored but rather that a considered decision using the results of all the available tests of normality needs to be made. For the Shapiro–Wilk and Kolmogorov–Smirnov tests, a P value less than 0.05 indicates that the distribution is significantly different from normal (Peat and Burton, 2005). For example, for the variable AGE, a low significance value, $p < 0.001$, for the two tests indicated that the distribution differed significantly from a normal distribution.

If a variable has a skewed distribution, it is possible to transform the variable to normality using a mathematical algorithm so that the outliers in the tail do not bias the summary statistics and p values, or the variable can be analyzed using non-parametric tests. For variables, like AGE, with non-normal distribution, transformations were done logarithmically during the analysis using parametric tests and, then, results were reverted to the original units for easier interpretation. Therefore, the variable AGE was transformed (logarithmic transformation) to normality. The advantage of logarithmic transformations is

that they give interpretable results after being back-transformed into original units (Bland and Altman, 1996). Other common transformations include square roots and reciprocals.

Descriptive/Summary statistics for each continous variable were produced. They included measures of central tendency and measures of dispersion (spread of the distribution) in the study sample. For continous variables whose distribution differs significantly from normal, e.g., AGE, the summary statistic used was the median and the corresponding interquartile range as opposed to the mean and standard deviation for the normally distributed variables. In a non-normal distribution, median and interquartile range provide a better estimates of central tendency than the mean and standard deviation since the median is more robust a measure of central tendency than the mean (Rosner, 2000), that is, it is not sensitive to departures from normality, e.g., in cases where the data are from a symmetric distribution with long tails or when the data have extreme values (Harris and Taylor, 2003).

Frequencies were also described in this data and by extension expressed as proportions. Peat and Burton (2005) warn of a common mistake in health studies involving frequencies and proportions whereby prevalence and incidence are used interchangeably although these terms have different meanings. Incidence describes the number of new cases with a condition divided by the population at risk. Prevalence is a term used to describe the total number of cases with a condition divided by the population at risk. The population at risk is the number of people during the specified time period who were susceptible to the

condition. The prevalence of an illness in a specified period is the number of incident cases in that period plus the previous prevalent cases and minus any deaths or remissions.

According to Gail *et al* (2010), to evaluate the extent to which an exposure is associated with a disease, e.g., participation in rice planting and schistosomiasis, we must often account or "control for" additional variables, such as age and/or sex, which are not of primary interest. Logistic regression is a mathematical modeling approach that can be used to describe the relationship of several independent variables to a dichotomous dependent variable, such as re-infection with *S. mansoni*. When a binary outcome variable is modeled using logistic regression, it is assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables (University of California, Los Angeles, 2011). The model is designed to describe a probability, which is always some number between 0 and 1. In epidemiologic terms, such a probability gives the risk of an individual getting a disease. The logistic model, therefore, is set up to ensure that whatever estimate of risk we get, it will always be some number between 0 and 1 (Gail *et al*, 2010).

After analysis was complete, analysis report was prepared as a joint effort between the research methods intern and the investigator whereby the investigator brought in the knowledge of the subject matter and the intern tied it to the statistical analysis outputs. This involved assistance in interpretation of statistical aspects such as the odds ratio and

adjusted odds ratio; for example, (AOR =2.51, 95% CI: 1.07-4.35) was reported as "a child from poor household was more likely to be re-infected with *S. mansoni* compared to one from a high economic status household (AOR =2.513, 95% CI: 1.07-4.35)". Alternatively, it could be reported as, "Re-infections with *S. mansoni* were 2.5 times more common in children from poor households as compared to those hailing from rich (high economic status households)".

Odds ratios have become widely used in medical reports and Bland and Altman (2000) put forward three reasons for this. Firstly, they provide an estimate (with confidence interval) for the relationship between two binary ("yes or no") variables. Secondly, they enable us to examine the effects of other variables on that relationship, using logistic regression. Thirdly, they have a special and very convenient interpretation in case-control studies. Despite their usefulness, odds ratios, Sackett, Deeks and Altman (1996), contend that they can cause difficulties in interpretation.

To gain more insight on the re-infection, follow-up studies which became evident in the course of statistical analysis were suggested as part of the research methodology support to this study. The suggestions included investigations into the determinants into the differential variations in re-infections in the three cohort schools and explanations as to what predisposes the children from poor households to re-infections (Is it poor nutrition

and hence poor immunity? Is it the differences in the sources of water used in those households or sanitation?).

The main challenge was that some categorizations of variables, e.g. economic status, which were found to be significantly associated with the disease, were not informative. This could be attributed to a flaw in the design of the study and the choice of the study sites. To be specific, the participating schools were assumed to be a 'cohort', that is they had a similar characteristics with respect to the disease in question, e.g., in terms of prevalence, re-infection rates. Analysis proved this assumption to be unfounded with the differences between the three schools being statistically significant with respect to the assumed 'cohort' attributes. As a recommendation, this could have been averted by conducting a baseline or a pilot study or approaching it as a comparative study rather than conducting a descriptive cross sectional study.

Other challenges encountered in the analysis were based on consolidating the theoretical RM knowledge gained in the course work and use it to do a comprehensive analysis. It was possible, for example, to generate the descriptive statistics but organizing and presenting the findings in a logical way still was a difficult task. Difficulties were experienced in reporting statistical outputs particularly on creating narratives that linked the statistical outputs with the subject matter of the study. These were solved by reading, widely, publications on the corresponding field of study and, even more important, consulting

senior statisticians at the institute. This shows there is a need for a research methods intern to work under the tutelage of an experienced statistician.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 Participating in research/consultancy

Successful resolution of the challenges posed by investigators will always require the close and honest collaboration of the two parties whereby the specialists brings in the statistical knowledge and the investigator brings in the knowledge on the subject-matter. The consultations usually involve choice of appropriate research design for a study, sample size determination, writing of technical passages that require statistical input and in interpretation of statistical outputs. Availability of a research methods consultant ensured that the research methodology challenges arising from research and research related activities were resolved hence enhancing the quality of research. Some of these consultations required revisiting the raw data and as such it is desirable for organizations to institute policies that ensure that the ownership of data from research belongs to the organization and/or the funding agencies as this ensures availability of such data sets in future even when the investigators leave the institutions.

5.2 Data management

Data are often expensive to collect. There is no point, therefore, in putting a lot of effort into making sure that the methods used for obtaining the data are of high scientific analytical quality if the same stringent quality controls are not maintained during data recording and computer entry. Data management can take the most time and, if not attended to carefully, can lead to errors in data recording, computer entry, storage and management that can spoil the subsequent statistical analysis. In data management, it is a very important step to separate the task of organizing the spreadsheets from that of actually entering the data. The organization of the spreadsheet in preparation for data entry should be informed by the envisioned statistical analyses. Spreadsheet such as Excel is very flexible and can be used for effective data entry, particularly for data sets with a simple structure. Their very flexibility, on the other hand, can result in poor data entry and management and hence a need for users to apply the same rigour and discipline that is obligatory with more structured data entry software. Features such as data validation, freezing windows, rows and columns and data auditing are available in spreadsheet packages if utilized can greatly enhance the quality of data. These features can also be augmented by other approaches such as double data entry. To promote use of these features and ultimately promote proper data management practices in-house trainings can be done to build the capacity of staff in the area.

5.3 Data analysis

The data analysis process is closely linked to data management and part of the analysis involves managing the data suit the analysis being done. The analysis should start with description of the study participants and move progressively to more complex analysis involving inferential statistics and modelling. Assistance in interpretation of the statistical outputs may be accorded to the investigator, for instance, by outlining the salient points that should be included in the write up.

To ensure accuracy and validity of statistical analysis outputs from one should pay attention to the following during data analysis;

- Ensure the sample is representative of the population of interest
- Understand the assumptions of the statistical procedures and ensure they are satisfied
- If multiple comparisons are made, replicate or use cross-validation to verify the results
- Keep in mind what you're trying to discover (the objectives)
- Look at the magnitudes of the study variables not just the p-values
- Once the analysis is complete, check if the study objectives have been attained

5.4 Recommendations

A research centre should have a research methods consultant or a statistician as this can greatly enhance the quality and quantity of research outputs. In-house trainings of scientists on proper data management practices should be done regularly and should focus on the features of common data management software(s) and how to utilize such to maximize data processing efficiency. After analysis is conducted, assistance in interpretation of the statistical outputs should be accorded to the investigator(s), for instance, by outlining the salient points to be included in the write up and on identifying gaps that call for further research and/or follow-up studies. Analysis, and hence the studies, should eventually pinpoint the specific issues that can be used to inform policies that will address the challenge(s) at hand. In the case of the analyzed study it was evident that further research needs to be done in order to inform the specific factors underlying the low economic status households being associated with the high re-infection rates.

REFERENCES

Academic Computing Services (2002). Excel Data Management. <www.albany.edu/ ~yhuang/excel-datamanagement.pdf>. [20 June, 2011]

Ader, J.H., Mellenberg, J.J. and Hand J.D. (2008). Advising on Research Methods: A Consultant's Companion, Johannes Van Kessel, Huizen, pp. 23, 40-43.

Alecxih, L., Corea, J. and Marker, D. (2011). Deriving State-Level Estimates from Three National Surveys: A Statistical Assessment and State Tabulations. Department of Health & Human Services/ASPE. <http://aspe.hhs.gov/health/reports/st_est/> [18 June, 2011]

Belle, V.G. (2008). Statistical Rules of the Thumb, 2nd ed. John Wiley & Sons, New Jersey, pp. 217-219, 272

Bland, J.M. and Altman, G.D. (2000). The Statistics Notes: The Odds Ratio. <www.bmj. com/content/320/7247/1468.1.full> [20 June, 2011]

Bland, J.M. and Altman, D.G. (1996). Statistics Notes: Transformations, Means, and Confidence Intervals. *BMJ*; **312**: 770. < www.bmj.com/content/312/7038/1079.full> [16 June, 2011]

Bowers, D. (2008). Medical Statistics from Scratch: An Introduction for Health Professionals, 2nd ed. John Wiley & Sons, West Sussex, p. 287

Chadha, V.K. (2006). Sample Size Determination in Health Studies. NTI Bulletin, 42/3&4, 55 – 62. < ntiindia.kar.nic.in/ntibulletin/...3.../03_MainPaper_Sampling.pdf > [18 June, 2011]

Chung, C. (2008). Logistic Regression with SPSS. < http://www.indiana.edu/~statmath /stat/all/cat/1b2.html > [10 June, 2011]

Coe, R. (2011).Which Test is Appropriate? <http://www.statistics-training.org/mod/forum /discuss.php?d=2556&parent=11591 > [18 June, 2011]

Eng, J. (2002). Sample Size Estimation: How Many Individuals Should Be Studied? <radiology.rsna.org/content/230/3/606.full.pdf > [18 June, 2011]

Feinstein, A. R. (2002). Principles of medical statistics. Chapman & Hall/CRC, Florida, pp. 42, 503

Fleiss, J. L. (1981). Statistical methods for rates and proportions, 2nd ed., Wiley, New York, p. 45

Frison, L. and Pocock, S. (1992). Repeated measurements in clinical trials: analysis using mean summary statistics and its implications for design. *Stat Med*; **11**:1685-1704.

Gail, M., Krickeberg, K., Samet, J.M., Tsiatis, A., Wong David, G. W., Klein K.M. (2010), Logistic Regression: A Self-Learning Text, 3rd Ed., Springer Science+Business Media LLC, New York, pp. 73-75, 209, 310-316

Harris, M. and Taylor, G. (2003). Medical Statistics Made Easy, Springer-Verlag, New York, pp. 45 - 46

Healy, M.J.R. (1993) Statistics from the inside: Data transformations. *Arch Dis Child*; **68**: 260–264

Hulley, S.B., Cummings, S.R., Browner, W.S., Grady, D., Hearst, N., and Newman, T.B. (2001). Designing Clinical Research: An Epidemiologic Approach, 2nd ed. Williams & Wilkins, Philadelphia, pp 65-84

Israel, D.G. (2009). Determining sample size. < http://edis.ifas.ufl.edu/pd006> [12 July 2011]

Lemeshow, S. (1990). Adequacy of sample sizes in health studies. Wiley & Sons, New York, p. 243

Lenth, V.R. (2001). Some Practical Guidelines for Effective Sample-Size Determination. < www.stat.uiowa.edu/techrep/tr303.pdf >. [19 June, 2011]

Levy, P.S. and Lemeshow, S. (2008). Sampling Populations: Methods and Applications. 4th ed., John Wiley & Sons, New York, p. 96

Mbinya, M. J. (2011). Factors contributing to re-infection with *Schistosoma mansoni* among primary school children: a case study of cohort schools in Mwea irrigation scheme central Kenya, Research proposal submitted in partial fulfillment for the degree of Masters of Science in Community Health and Development, of the Great Lakes University of Kisumu

McDonald, J.H. (2009). Handbook of Biological Statistics, 2nd ed., pp 112-117. Sparky House Publishing, Baltimore. <udel.edu/~mcdonald/statconf.html> [18 June, 2011]

McGinn,T. (2004). RHRC Consortium Monitoring and Evaluation ToolKit: PPS Sampling Technique. < www.rhrc.org/.../55b%20PPS%20sampling%20technique.doc > [18 June, 2011]

Muraya, K.P. and Chege, G.W. in Muir-Leresche, K. and Coe, R. and Ekwamu, A. (2009). GEAR: Graduate Environmental and Agricultural Research: A Guide to Effective and Relevant Graduate Research in Africa. RUFORUM, Kampala, pp. 194-205

Mwobobia, I. K., Ng'ang'a, P.M., Muniu, E.M., Mukoko, D.A. and Njenga, S.M. (2011), Insecticide treated nets based malaria control in Kenya: a micro-level coverage and use in a district under national malaria control programme. Proceedings of the 2011 Joint Research Conference: (4th Walter Sisulu University International Research Conference, 8th Society for Free Radical Research-Africa (SFRR), 31st African Health Sciences Congress (31st AHSC) & 4th International Conference of the Promotion of Traditional Medicines (PROMETRA)), East London, South Africa

Mwobobia, I., Mutua, A., Musyoki, S., Ng'ang'a, P., Muniu, E., Kihara, J., Njenga, S. (2011). Proposal: A comparison of the impact of de-worming with that of integrated de-worming and Community-Led Total Sanitation on schistosomiasis and soil transmitted helminths prevalence and re-infection rates in coastal Kenya

Mwobobia, I. K., Ng'ang'a, P.M., Muniu, E.M., Mukoko, D.A. and Njenga, S.M. (2011), Evaluation of Sanitation Hygiene and Insecticide-Treated Nets Use in Mwaluphamba Location, Proceedings of the 1st KEMRI Annual Scientific and Health (KASH) Conference, Nairobi, Kenya, p. 39

Ospina,D. and Ortiz, E.J. (2001). Statistical Research and Consulting in Universities from Developing Countries: The Colombian Case. < http://www.stat.auckland.ac.nz/~iase/ publications /9/295.pdf > [8 July 2011]

Patel, B.K., Muir-Leresche, K., Coe, R. and Hainsworth, S.D. (2004). The Green Book: A Guide to Effective Graduate Research in African Agriculture, Environment, and Rural Development. The African Crop Science Society, Kampala, p. 248.

Peat, J. K, and Barton, B. (2005). Medical statistics: A Guide to Data Analysis and Critical Appraisal, 1st ed., Blackwell, Oxford, pp. 9-12, 219 – 221

Richardson-Kageler, S., in Muir-Leresche, K. and Coe, R. and Ekwamu, A. (2009). GEAR, Graduate Environmental and Agricultural Research: A Guide to Effective and Relevant Graduate Research in Africa. RUFORUM, Kampala, pp. 208-218

Rosner, B. (2000). Fundamentals of Biostatistics, 5th ed. Pacific Inc, Duxbury, pp. 308-311

Rowe, K. A., Lama, M., Onikpo, F. and Deming S.M. (2001).Design Effects and Intraclass Correlation Coeffecients from a Health Facility Cluster Survey in Benin. *International journal for quality in health care*, **14**(6), 521

Sackett, D. L., Deeks J.J. and Altman D.G. (1996). Down with odds ratios! Evidence-Based Med; 1: 164-6. < http://www.bmj.com/content/317/7168/1318.1.full>. [20 June, 2011]

Schoenbach, J.V. (2000). Data management and data Analysis. <http://www.epidemiolog. net /evolving/DataManagement.pdf>. [18 June, 2011]

Schwalbe, K. (2009). Information Technology Project Management, 6th ed. Cengage Learning, Boston, p. 399

Shia, C.B. (2001). How to Think About Statistical Consultation? Learning from Data. <http://www.stat.auckland.ac.nz/~iase/publications/9/371.pdf>. [20 December 2011]

Statistical Services Centre (1998). Data Management guidelines for experimental projects, Reading

Statistical Services Centre (2001). Approaches to the Analysis of Survey Data, Reading

Statistical Services Centre (2006). Confidence and Significance: Key Concepts of Inferential Statistics, Reading.

Stevens J. (1996). Applied multivariate statistics for the social sciences, 3rd ed. Lawrence Erlbaum Associate, New Jersey, pp. 237–260

University of California, Los Angeles (2011). How do I interpret odds ratios in logistic regression? < www.ats.ucla.edu/stat/stata/faq/oratio.htm > [20 June, 2011]

US Census Bureau (2007). Technical Paper 63: Current Population Survey - Design and Methodology, pp 4-8. <http://www.bls.census.gov/cps/tp/tp63.htm> [18 June, 2011]