An Application of SAS® in a Large Scale, Foreign Epidemiological Study

Dwayne D. Oland, M.S.¹ and Sally K. Stansfield, M.D.²

¹U.S. Army Medical Research Institute of Infectious Diseases
Fort Detrick, Frederick, MD 21701-5011, USA
²World Health Organization, Geneva, Switzerland

Abstract

The SAS system was used in a twelve year population registration-based epidemiological study conducted by the Hospital Albert Schweitzer in the rural Artibonite Valley of Haiti. The study incorporated vital statistics and health status data obtained during home visits by community health workers and mortality and morbidity records obtained from the hospital admission and outpatient records. The epidemiological analysis included the first detailed census of the area, disease morbidity and mortality rates, and the effect of socioeconomic, health program, and geographic factors on disease rates.

Base SAS and SAS/GRAPH® provided the full spectrum of tools, procedures, and techniques required for effective data management and analysis in this study.

Background

Since opening in 1956, the Hospital Albert Schweitzer (HAS) has provided primary health care services for residents of the Artibonite Valley in rural Haiti. The population of the 380-Km² district served by HAS numbers over 165,000. The 150-bed hospital tallies 3500 admissions and provides 50,000 patient days of inpatient care annually. Vital statistics are recorded by HAS-supported community health workers (CHW) during domiciliary visits, and records are kept of both inpatient and outpatient diagnoses. This excellent health information system provides an unusual opportunity to describe the health status of a population in this poorest nation in the western hemisphere.

Data Collection

Vital statistics and health status data were recorded during home visits by CHWs for each member of every household in the HAS medical district. Records of these visits were abstracted and coded, including a house identifier and data on age, sex, hospital number, and history of pregnancy or illness. Vaccination status and nutritional status were generally recorded for children under five. The roof type (i.e., straw, tile, tin, or concrete) and condition of the home were recorded for each household as a rough indicator of socioeconomic status. The total number and dates of CHW visits were also coded for purposes of this analysis. Additional data entered for each household included the identity and duration of service provided by their CHW, the name, distance from household, and duration of service provided by the nearest dispensary, and the geographic location of each village (i.e., mountain or valley). A total of 190,862 CHW records were coded for this study.

Disease-specific morbidity and mortality data were obtained by review of hospital admission and outpatient clinic records over 12 years. Information coded and entered from the hospital and clinic records included a diagnosis code for the selected illnesses, hospital number, location of residence, date of admission or clinic visit, number of days of inpatient care, and outcome (i.e., death or discharge). A total of 98,741 hospital and clinic records were coded for this study.

All data were coded on columnar data sheets and sent to the U.S. Army Medical Research Institute of Infectious Diseases for data entry, data verification, and analysis. Data entry was done under contract by firms specializing in international work.

Data Management

All data management and analysis were performed on an AMDAHL model 470/V7 with base SAS and SAS/GRAPH software. Two SAS data sets were established from the domiciliary visits and hospital-based information. Data verification and quality control were conducted in three separate phases to ensure data integrity:

1. Histograms and charts created by PROC CHART and PROC FREQ were used to extract outliers and suspect values for each SAS data field. These results provided a review of obvious errors by data coders and misinterpretations by the data-entry personnel. This technique was also used to detect redundancy in unique fields (i.e., multiple recording of the same household or the same patient visit) and improbable situations (i.e., households with 232 occupants).

2. Unresolved suspect values from phase 1 were tabulated in various ways with PROC TABULATE. These tables combined several SAS fields that were logically related in order to detect patterns of errors. These tables were useful in detecting systemic deviations from the coding protocol and specific idiosyncrasies of the HAS data encoders and community health workers.
3. Further analysis by the above techniques were used to evaluate medical impossibilities (i.e., pregnant males). Tables of multiple diagnoses and age and sex specific conditions were reviewed by medical personnel to determine suspect data.

All suspect data were reviewed by the database manager and the principal investigator and submitted to HAS for reconciliation. Changes were made to the database as required and any suspect values that could not be resolved were designated as missing values. After all changes were made, the entire verification process was repeated to ensure data integrity.

Data Analysis

The domiciliary visit and hospital morbidity and mortality data sets were analyzed by a wide variety of SAS statistical and graphical techniques and procedures. The epidemiological analysis was performed in three distinct phases, as shown on Figure 1.

1. Population Census

Because no dependable census information was available for this region, it was necessary to tally census information by geographical region, village, age, and sex over the 12 years of the study. A series of interrelated computations by PROC MEAN were used to derive population figures from household size, birth, death, immigration, and emigration data in the domiciliary visit data set. The population was also characterized by several socioeconomic factors, such as condition of the home, household size, and spacing and number of children.

2. Overall Disease Morbidity and Mortality

The hospital morbidity and mortality data set was analyzed to show overall morbidity and mortality rates of 33 diseases and medical conditions found in this region. Disease-specific rates were examined over time and differentiated by sex and age.

3. Effect of Socioeconomic, Health Program, and Geographical Factors

The domiciliary visit and hospital mortality and morbidity data sets were merged for this phase of the analysis, by hospital numbers or geographical areas. Disease rates were analyzed in light of factors such as condition of the home, household size, spacing and number of children, immunization status, number and frequency of community health worker visits, distance from dispensaries, and terrain (valley or mountain). The effectiveness of preventative health care services including health and nutrition education, supplementary food distribution, tuberculosis screening, immunization against tetanus, diphtheria, pertussis, polio and tuberculosis as well as oral rehydration therapy were examined. The effectiveness of distributed health care in the form of remote dispensaries and community health workers was also examined by evaluating mortality and morbidity rates before and after intervention by health care professionals.

**Figure 1**
Results

The effectiveness of SAS as a tool in the overall management and analysis of large-scale, international, epidemiological studies is best illustrated by the final report issued by HAS. The 280-page report, which was the first long-term, population registration-based epidemiological study in this area, contained over 120 original charts, 8 histograms or graphs, and 65 tables from base SAS and SAS/GRAPH procedures characterizing the epidemiology of the Artibonite Valley. Most of these published materials were printed on line-or letter-quality printers or drawn by plotters. During the course of this study, no other statistical, graphics, or database software was used. The SAS system provided the full spectrum of tools, procedures and techniques required for effective data management and analysis.

Conclusion

Base SAS and SAS/GRAPH were effective tools in the management and analysis of this large, foreign, epidemiological study. However, centralized management and analysis of epidemiological data away from the foreign study site involves several risks. Data entry and verification can become tedious and time consuming and many times the data integrity is compromised because they cannot be effectively or easily reviewed by the foreign research staff.

The introduction of SAS and SAS/GRAPH for personal computers coupled with advances in microcomputer technology, allow researchers to distribute the data entry, verification and management functions to the remote site (Figure 2). An automated data management system designed for remote medical studies has been used effectively in Argentina and China and uses portable microcomputers and optical scanners. Data entry is accomplished through the use of mark sense forms and verification and management is performed by using DBASEII+ and SAS software. The use of mark sense forms or laptop and hand-held computers greatly facilitates the data-gathering effort by prompting the interviewer with acceptable codes and documentation and allowing data verification and editing at the time of the initial interview.

After this data is transferred to the foreign hospital, clinic or study site, the foreign research staff members can monitor the progress of the study, check for inconsistent data and perform preliminary analyses on the data. Using SAS and SAS/GRAPH software, these preliminary analyses might suggest refinements in the study which would enhance the overall success of the project. The centralized analysis site would receive verified and error-free SAS data sets in a machine-readable format which would allow the overall analysis to proceed at a faster rate.

The use of SAS tools and products at each stage of this process could greatly enhance the data management and reduce the time for analysis in large-scale epidemiological studies.

DECENTRALIZED DATA MANAGEMENT

<table>
<thead>
<tr>
<th>Location</th>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical mark sense forms</td>
<td>- Entry level editing and prompting</td>
</tr>
<tr>
<td>Lap top computer</td>
<td>- Elimination of extraneous codes and suspect data</td>
</tr>
<tr>
<td>Hand held computer</td>
<td>- Checks for inconsistent data</td>
</tr>
<tr>
<td>Microcomputer</td>
<td>- Daily monitoring of health workers data</td>
</tr>
<tr>
<td>Centralized analysis site</td>
<td>- Preliminary analysis of data</td>
</tr>
<tr>
<td>Minicomputer &amp; mainframe</td>
<td>- Data arrives in machine readable format</td>
</tr>
<tr>
<td></td>
<td>- Data is verified and error free</td>
</tr>
<tr>
<td></td>
<td>- Analysis can proceed at faster rate</td>
</tr>
</tbody>
</table>

Figure 2
Acknowledgments

The authors express their appreciation to their colleagues at Hospital Albert Schweitzer whose dedicated efforts have made this work possible.

References


2. Bertrand, W. "Use of Microcomputers in Health and Social Service Applications in Developing Nations; CRO Critical Reviews in Medical Informatics, Volume I, Issue 3.


The views of the author(s) do not purport to reflect the positions of the Department of the Army or the Department of Defense.

Approved for public release, distribution unlimited.

SAS and SAS/GRAPH are registered trademarks of the SAS Institute, Inc., Cary, N.C., USA.

Dwayne D. Oland
Applications Development
Computer Science Office
USAMRIID
Fort Detrick, MD 21701-5011
(301) 663-7514