University of Leeds

University Computing Service

SAS Institute Funded Project on Epidemiological Computing

Progress Report No. 4

30th October 1997

Contents

1. Introduction

Previous reports have concentrated on summarising progress made since the preceding report. The purpose of this report is three-fold. In addition to reporting on developments since report No. 3 was issued earlier this year, the report provides a review of the project from its beginning, summarising the main activities and achievements to date, and goes on to discuss the prospects for future development. In the interests of providing a complete and balanced view, some material has been included which has already appeared in earlier reports.

2. Project Aims Re-Visited

The aims of the project were set out clearly in report No. 1.

The primary aim was to build an application which will allow the LRF to produce tables, analyses and graphs required for the leukaemia disease atlas. A secondary aim was to be able to extend the applicability of the system to other diseases studies. In addition, publicity was recognised as an important part of the project.

This project involves more than building an application. The LRF have been engaged for more than ten years in the investigation of disease incidence and have been successful in the production of the disease atlas of leukaemias. In building this application we are attempting to bring unity to much of the computing carried out by the LRF in the analysis of disease incidence. The new application will allow new ways of working and should also secure significant productivity gains. If successful, the influence of their work in the field of epidemiology has the potential to interest other research groups in using the SAS system for their work.

This report relates the work carried out so far to meet these expressed aims.

3. The Learning Curve

3.1 Understanding the Users' Requirements

The first task for the team was to understand the users' needs and to draw up a requirements document for the project. To that end, regular meetings were held with the users over a period of about four months culminating in the production of a *Specification of Requirements.*

3.2 Learning and Training

In parallel with the efforts to specify the Users' requirements, an education programme was set up to provide members of the team with knowledge and skills needed for the project and to bridge the gap of understanding between developers and users. The needs were three-fold:

- SAS skills
- Familiarity with epidemiological methods

- Understanding of relevant statistical methods

Endeavours undertaken include:

- Attendance at SAS Training Courses
  - Introductory Course on SAS at Manchester University (DA)
  - Courses 1-3 of SAS/TUTOR (DA)
  - Applications Development Using the SAS System(DA)
  - Object Oriented Programming with the SAS System (DA,PN)

- Learning SAS/AF and SCL (DA,PN)

- Learning Statistical Analysis (DA) using:
  - Tuition from PN
  - 'Statistics for the Terrified'

- SAS Training for the LRF via:
  - A 2 day 'Introduction to SAS' course provided by PN
  - A report, produced by PN, summarising the statistical procedures available in SAS relevant to epidemiology. A diskette was also provided containing a set of sample programs illustrating the use of a number of these procedures.

- Learning SAS/GIS via the GIS tutorial (DA,PN)

- Learning the basics of epidemiology (DA,PN)

3.3 Problems with Spatial Data

Numerous difficulties were experienced in attempting to import the digitised map boundaries of England and Wales from UKBORDERS at Edinburgh. Problems were encountered with both the data and with SAS/GIS. Some time was lost as a result of these problems, none of which were of local origin. However, the maps now available look satisfactory and should enable developments on the GIS front to proceed normally. We are grateful to SAS Institute for their technical expertise and other support in this area. The provision of a utility to convert GIS map data sets to SAS/GRAPH map data sets, which Steve Morton has developed, will also contribute to the generation of static (choropleth) maps.

4. Data Modelling for an Epidemiological Data Warehouse

The data required to compute tables and maps of disease incidence rates are diverse. In addition, a variety of problems inherent in the data supply necessitated careful planning in the design of a suitable data model. For example, some of the counties contributing to the collection of case data have districts which are not included in the study. This raises a practical difficulty in computing incidence rates since county level population data can not be used directly. Instead, it is necessary to obtain an adjusted population figure by obtaining district level population data and aggregating over just those districts that are contributing cases to the county in question.

After much thought and discussion with the LRF a data model was adopted, founded on relational database theory but extended to include summary tables at all levels in the area hierarchy.

Full details of the problems posed by the data and the eventual model adopted can be found in the paper by Allon and Nicholson (1997).

5. Analysis

5.1 Integration of Existing Analyses

The challenge in the area of analysis was to take a wide range of analyses currently implemented using a variety of languages and packages, and each with its own input and output formats, and to implement them using the SAS system.

With exception of one program, conversion has been achieved. In addition, the SAS implementations communicate via a common file format, the SAS data set. This facilitates a smooth workflow and eliminates the inefficiencies arising from the use of non-standard output formats.

For a full list of the analysis procedures currently implemented, see Section 7.3 below.

5.2 Computationally Intensive Procedures

A translation of a C program for the Maximum *a-posteriori* estimation (Mapest) model was performed. The program includes several routines which perform specific numerical operations, including matrix decomposition, computation of determinants, quadratic interpolation and maximisation of a likelihood function using iteratively re-weighted least squares and considerable effort was required to translate the routines into SAS. The majority of the numerical algorithms were coded as SAS/IML modules.

The results, however, were not encouraging. Whilst perfect agreement was reached with results obtained by the C program applied to 58 age-groups, the performance of the SAS conversion was extremely poor. A number of alternative approaches were considered. Details of these are set out in Report No. 3.

After consideration of these alternatives, it was felt that the most practicable solution would be to exploit either SAS/TOOLKIT or the MODULE facility to harness the power provided by the C language. Accordingly, this remains high on the list of priorities for future developments.

6. Publicity

6.1 Contributions to SEUGI 15

The following paper and poster were presented at the SEUGI Conference in Madrid in May 1997.

(i) Paper: Analysing the Incidence of Leukaemia in England and Wales

Author: Paul Nicholson
Co-authors: Deborah Allon, Richard McNally, David Rowland

(ii) Poster: Data Modelling for an Epidemiological Database

Author: Deborah Allon
Co-Author: Paul Nicholson

The paper provides an overall description of the application being developed. The poster discusses alternative data models and describes the data model ultimately adopted.

Inevitably, the preparation of contributions to SEUGI detracted from application development for a substantial period of time during November through to March. Each article submitted for publication in the conference proceedings ran to more than 4000 words and included numerous figures and graphic images. In addition, further effort was required to produce the presentation form of each article - a Powerpoint presentation consisting of 30 slides for the paper and 10 heat-sealed and mounted A3 displays for the poster. However, it was recognised from the start, as expressed in the 'Project Aims' in Report No. 1, that attention to publicity would form an integral part of the project. We are therefore pleased to report that the paper won the award of 'Best Paper' in the Statistics stream.

6.2 Seminars and Newsletters

On return from SEUGI, and following further development work on the application, a presentation similar to that presented at SEUGI was given at the University of Leeds. The seminar was attended by users from a variety of departments within the University, including several users from medical departments. A précis of the published paper was also published in the Computing Service Newsletter. An updated version of that article was subsequently made available to SAS Institute (UK) for publication in a forthcoming edition of the InSite Newsletter, edited by Bruce Bovill of SAS Institute (UK).

6.3 Contacts with the Research Community

6.3.1 Visit by Professor Theodore Holford

On January 22nd, 1997, we were pleased to receive a visit from Theodore Holford, Professor of Epidemiology at Yale University in the USA. Professor Holford was on a study visit to the UK and visited Leeds to present a talk on the cohort analysis of cancer data to the Medical Statistics Workshop within the University and to visit the Leukaemia Research Fund. We were able to demonstrate the application under development and also to

demonstrate the use of SAS to analyse data pertaining to the incidence of prostate cancer taken from one of his own research papers. Professor Holford currently uses a batch style of working in his own computing, making manual changes to existing programs, and was attracted to the ability to specify requirements interactively without the need for programming ability. He also expressed interest in the use of GIS for the display of results. We agreed to keep him informed of our progress on the project.

6.3.2 Paediatric Epidemiology Group, University of Leeds

Discussions have taken place with Dr. Tricia McKinney of the Paediatric Epidemiology Group at the University of Leeds concerning the scope for the use of the application under construction in the work of her group. Dr. McKinney was formerly employed within the LRF centre at Leeds and participated in the early work in establishing the LRF Disease Atlas. She therefore understands the nature of the work that the LRF are engaged in and their computational requirements. Dr. McKinney has identified a number of possibilities for the application or extension of the system under construction to her work. These are described in Section 8 on Future Developments below.

6.3.3 The Dutch Connection

Jan Willem Coebergh, an epidemiologist based in the Netherlands and involved with the Eindhoven Cancer Registry, has expressed an interest in using the application in his work. Jan has collaborated with Professor Ray Cartwright from the LRF on the European study of lymphomas and has maintained links with the LRF at Leeds.

7. Current Project Status

A substantial amount of effort has been expended since the third report was released in February. A not insignificant part of that effort went into producing two submissions for SEUGI as described earlier. Since the return from SEUGI in May, the thrust of development has been on the integration of analysis modules into the application and the development of the user interface leading to the production a prototype for the application.

7.1 Development Platform

Immediately on return from SEUGI, development was transferred to Windows NT. This move was prompted after a number of errors had been encountered using some of the new objects provided in SAS/AF under Windows 3.11. The use of SAS 6.12 under Windows NT has proven to be significantly more reliable than Windows 3.11.

7.2 User Interfaces

The availability of the new TAB object in release 6.12 of SAS prompted a re-think in the design of the user interface. Formerly, we had been using a hierarchical sequence of SAS/AF FRAME screens for the part of the application concerned with the specification of data required for analysis. The use of a Tab Layout allows the variety of screens required for data selection to be organised within a single frame.

Figure 1 shows the Tab Layout object used to facilitate disease group selection. The available set of disease groups is listed in alphabetical order in a list box with their full descriptions alongside. Linked to this listbox is an organisational chart which displays the hierarchical structure of the disease groups graphically. The user is free to select the required group from either the listbox or the chart. The value selected is automatically communicated to the other object.

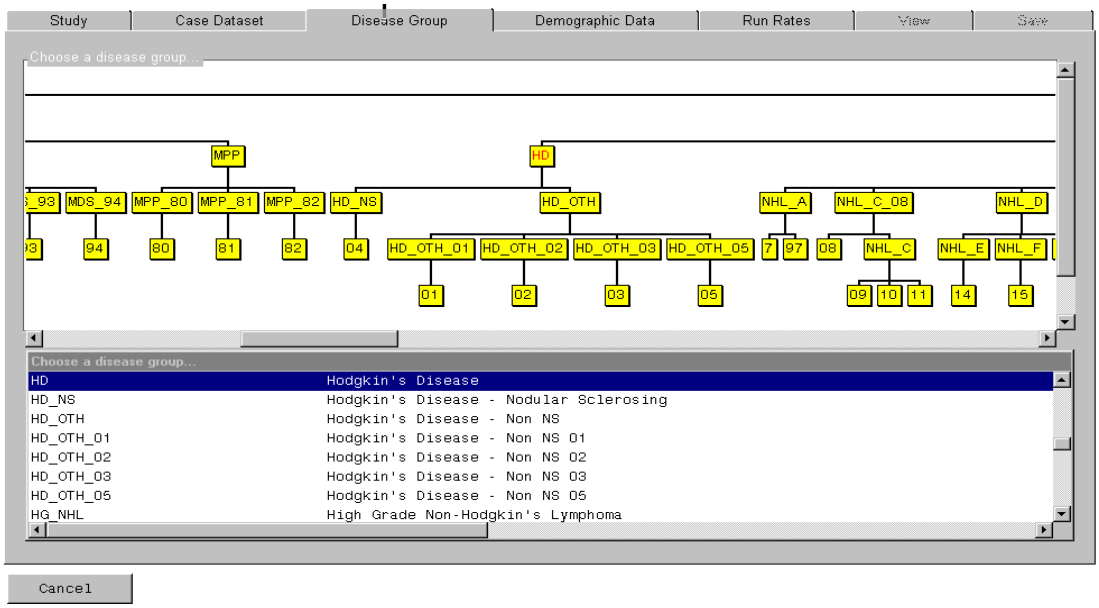

Figure 1. Disease Group Selection Screen

Figure 2 shows the screen used to specify the demographic data selection criteria.

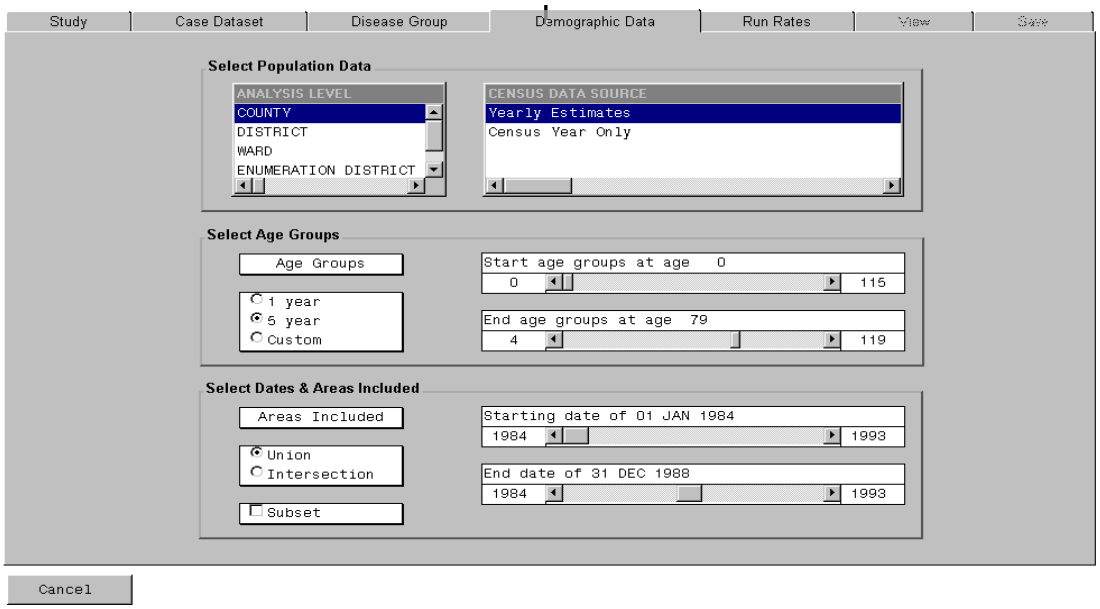

Figure 2. Demographic Data Selection Screen

SAS 6.12 also saw the emergence of the MAP object which provides a FRAME interface to the SAS/GRAPH procedure GMAP used for choropleth maps. If a geographical subset of data is requested, a separate screen is displayed which uses a MAP object to display a map inviting the user to select the area required for analysis. This is illustrated in Figure 3.

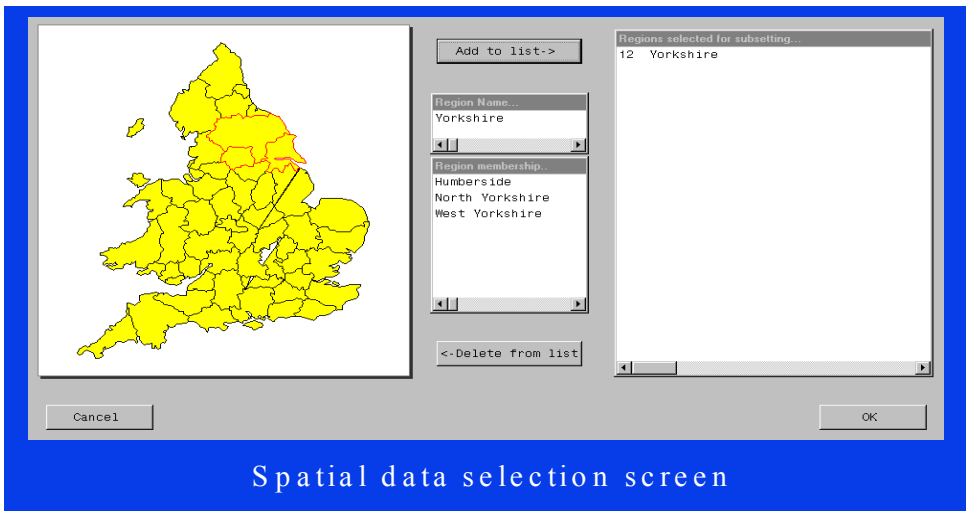

Figure 3. Spatial Data Selection Screen

The Tab Layout object inherently provides a summary of the choices made and gives the user the chance to change any of them if necessary. Another big advantage of the TAB object, over the previous method of a sequence of frames, is the ease with which further tables can be produced. For example, it may be required to produce a table for a new disease group, keeping all other factors the same. All that is required is to click on the disease group tab and run the rates calculation again.

A special interface was also required to facilitate the definition of an arbitrary number of age-bands. These are required by the Breslow-Day age-standardisation process which requires analysis to be carried out in age-bands in order that tests of heterogeneity can be performed.
This is depicted below in figure 4. Currently, the age-bands must be contiguous. The ability to define non-contiguous age-bands is planned.

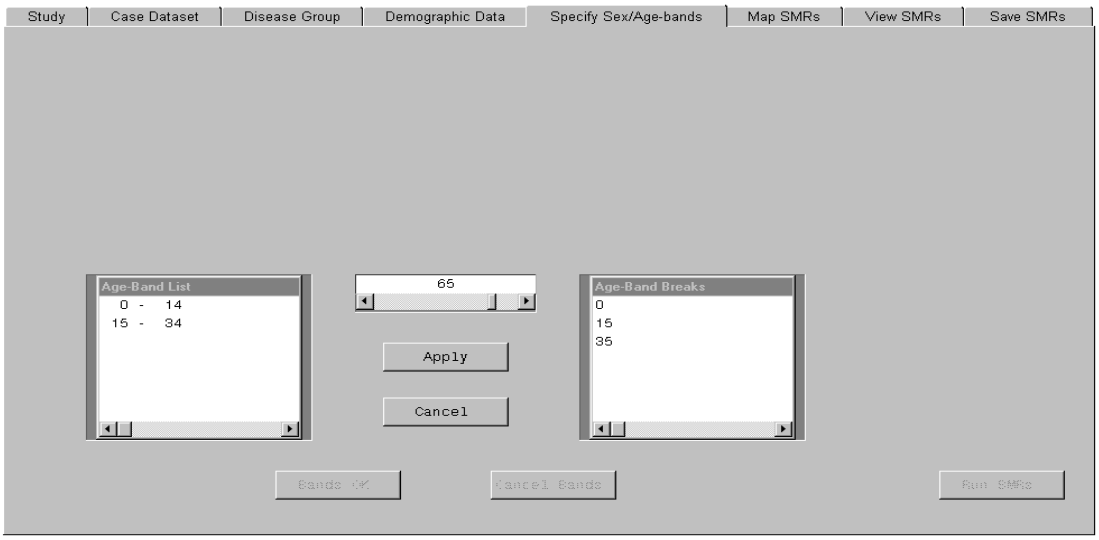

Figure 4. Age-band Definition Screen

7.3 Functionality

The following procedures have now been integrated into the application:

- Calculation of age-specific incidence rates
- Direct age-standardisation
  - Standards include World, African, European, Uniform, World-truncated
- Breslow-Day indirect age-standardisation
  - Analysis in age-bands or unbanded
- Poisson regression modelling of SMRs including:
  - Confidence Intervals
  - Tests of Heterogeneity
  - Lancaster's two-sided mid-P values
  - Breslow's test for over-dispersion
- Empirical Bayes estimates of SMRs
  - 'Shrinkage' estimates based upon a Gamma prior distribution
  - 'Shrinkage' estimates based upon a Log-Normal prior distribution
- Choropleth maps of SMRs
- Choropleth maps of smoothed SMRs (SRRs)

The majority of the current requirements for graphical display can be satisfied using SAS/GRAPH. The new MAP object provided in SAS 6.12 has been used for the production of static choropleth maps and the production of age-specific incidence curves (currently under implementation) is catered for by the GPLOT procedure.

7.4 Names

A fundamental problem facing the application developer is the choice of a workable system for the storage and naming of the various objects created by the application. In this project, this includes rates tables, SMR tables and graphs. In the application currently used by the LRF, very long names, constructed by concatenating the values of various data selection factors, are used. Under NT, long file names could be used for external files. However, portability would be lost by adopting this approach. Furthermore, many names used within the application are restricted to a maximum of eight characters (SAS data set names for example).

An obvious approach to breaking the complexity of this problem is to use a sub-directory structure which maps onto the various levels of the data selection factors. This is the approach taken in this application. The name of the object saved specifies what it is and the path to the object specifies how it was created. A benefit of this approach is that the user does not have to wrestle with long names. Data objects can be located either by specifying factor values via the user interface provided or by using the operating system's file manager to locate the object in the directory hierarchy.

Methods have also been written to save and locate the various object types depending on the criteria selected. These will be useful when developing the batch capability in the future.

7.5 Application Prototype

The functionality described in 7.3, above, represents a significant proportion of the analysis and reporting features specified in the original requirements. Taken together, they will allow the LRF to carry out geographical analysis of the variation in disease incidence for a variety of diseases or disease groups and age-groups, broken down by sex at both county and district level. Although some further work is required on mechanisms for saving and printing results and also for saving settings, the application should soon be ready for testing by the end-users.

8. Future Developments

Three main future development strands for the project are proposed:

- Implement those features outstanding from the original specification
- Extend the application to other disease studies
- Extend the scope of the application by including new functionality

In addition the scope for the application of SAS on other projects is discussed.

8.1 Features Outstanding

A number of features from the original Specification of Requirements remain to be implemented. They fall into four main areas: Analysis and Graphics, Data Manipulation, GIS and General Applications.

Analysis and Graphics

- Empirical Bayes smoothing using Maximum A-Posteriori (MAPEST)
  - this requires the ability to execute external C code (either using the MODULE function or SAS/TOOLKIT)
- The graphical display of smoothed age-specific incidence curves
- Analysis of secular trends in SIRs using linear regression and Poisson regression for multiple age groups
- The graphical display of secular trends in SIRs
- Seasonal models for SIRs
- Analysis of clusters using a variety of methods including:
  - Potthoff-Whittinghill
  - Besag-Newell
  - NNA

Data Manipulation

- Facility to handle a variety of age-groups
- Facility to handle a variety of age-bands (in indirect standardisation)
- Data selection by varying criteria (links with GIS use)
- Data editing (disease codes, biopsy codes etc.)
- Estimation of yearly data using interpolation

GIS

- Definition of areas in a variety of ways including:
  - urban/rural classification
  - deprivation index (e.g. Thomson)
  - by geographic feature (e.g. coastal, estuaries, high bracken levels)
  - areas within a specified distance of a point source
  - areas within a specified distance of a line (e.g. power lines, rivers, roads)

- Analysis of areas that satisfy a given selection criterion
  - and analysis of areas excluded by the same criterion
  (N. B. This is relevant to post-hoc cluster analysis described in 8.3.1 below)

General Application

- Link to enable C codes to execute
- Batch facilities
  - in the first instance, the ability to save a 'script' to serve as a template for similar runs in the future would be useful
  - in the longer term, a true batch facility to allow a run which produces high-volume output to be deferred for non-interactive execution is desirable
- Use of a WWW intranet for dissemination of selected results

8.2 Application to Other Disease Studies

As indicated earlier, in **6.3.2**, discussions have taken place with Dr. Tricia McKinney of the Paediatric Epidemiology Group at the University of Leeds concerning the possible application of the system to other disease studies.

Dr. McKinney has identified a number of areas of interest to her group. Some of these could be satisfied by application of the system currently specified. Others would require an extension of functionality. They include:

- The Yorkshire Childhood Diabetes registry
- The Childhood Cancer Register
- Analysis of clusters including:
  - Knox test (using Mantel modification)
  - Potthoff-Whittinghill test

Regarding the existing registries, some work would be required to transfer data currently resident in an Access database to the SAS system. However, the nature of the data is very similar to that used by the LRF and the prospects for a successful transfer look promising.

8.3 Extensions

8.3.1 Post-Hoc Cluster Analysis

A cluster is an aggregation of cases in terms of a disease group, time and space where the number of cases is statistically greater than one would expect when the natural history of the disease and chance fluctuations are taken into account.

A post hoc cluster is the most common circumstance likely to require investigation. It initially relies on observations of past events - almost any grouping of cases of disease in time and/or space - and for this reason it is almost beyond definition at the outset of an investigation. Further, it can be any close aggregation of cases deemed by the observer to be remarkable, unusual, sinister or for other reasons worthy of investigation. Specific case associations and often causes tend to be inferred in the early reporting.

The investigation of a post-hoc cluster requires the calculation of observed and expected numbers of cases in small areas as well as associated Poisson probabilities. The implementation in SAS could involve use of its GIS capabilities. It has relevance to the LRF in the investigation of reported “clusters” following public health concern. Others interested in investigating post-hoc clusters include academic departments of public health and public health consultants/departments both in the UK and world-wide.

A handbook and guide to the investigation of clusters of diseases (Arrundale, et al 1997), compiled at the LRF in Leeds, has been supplied to all district health authorities in the UK. The availability of an application to conduct the analyses outlined in that text would be of interest to those authorities.

8.3.2 Disease Surveillance

This involves the continuous monitoring of an area to look for an excess of cases. It uses both the CUSUM and SETS methods. Whilst it is not so relevant to the LRF, it could be of wide use to public health departments and other disease registers. In the UK the best example is probably the ONS congenital anomalies scheme set up as a result of the thalidomide tragedy.

Ref.: Arrundale et al (1997).

8.4 New Projects

Discussions with staff from the LRF and Dr. Tricia McKinney have identified a number of other projects which may also be suited to the SAS system.

8.4.1 National Childhood Cancer Study

A unique nationwide investigation into the causes of cancer in children, commenced in April 1992, has now amassed data on 1000 children diagnosed with cancer every year in England and Wales together with data on twice as many comparable healthy children. Responsibility for the study is in the hands of a management committee under the chairmanship of Sir Richard Doll and consists of clinicians, epidemiologists, biologists and a representative from the National Radiological Protection Board. All results will be stored centrally at the LRF at the University of Leeds for analysis and interpretation by the management committee.

8.4.2 ONS Occupational Health Study

Studies of the association between occupation and disease incidence, based upon routinely collected data, have made use of empirical Bayes estimation procedures to estimate the ratios between observed and expected number of cases. A graphical technique (probability plotting) has been used in which anomalous associations are revealed as outliers.

Ref.: Carpenter et al (1997).

9. Summary

Following a prolonged initial learning phase, made more difficult by technical problems beyond our control affecting the supply of spatial data, a solid foundation for the application was established in the form of an epidemiological data warehouse. The model upon which the warehouse is based has been carefully chosen to be able to address a range of problems inherent in epidemiological data of the kind handled by the LRF.

In parallel with the development of the data warehouse, work proceeded on the development of analysis modules. With the exception of one computationally intensive algorithm, a wide variety of programs written in FORTRAN, C and GENSTAT have been successfully implemented in SAS. These programs now communicate via SAS data sets facilitating a smooth workflow in the application. The use of the MODULE facility (or possibly SAS/TOOLKIT) is planned to facilitate incorporation of the outstanding computationally intensive codes.

A break in development was taken to prepare submissions for SEUGI 15 in Madrid. Though frustrating in its effect on productivity, this provided valuable publicity for our work. It was also rewarded with the prize for 'Best Paper' (Statistics).

The application is now close to being a prototype suitable for experimental use by the end user, although some further work is required on the provision of commonly required mechanisms, such as mechanisms for the storage and printing of results and saving preferences. Also, some documentation and training will be required when the prototype is handed over to the end-user.

Directions and priorities for further development have already been identified. In the short term, the emphasis will be on completing the analysis and graphics facilities cited in the original *Specification of Requirements*. Once those are in place, attention will turn to additional statistical requirements, to the development of the use of SAS/GIS and to the development of data management capabilities. In parallel with the development of the application, it is hoped that contact with other research groups will afford the opportunity to use the application on other disease studies. This should provide a further valuable test of the usefulness of the application for descriptive epidemiological studies in general.

References

1. Allon, D. and Nicholson, P. , Data Modelling for an Epidemiological Database. SEUGI 15 Conference Proceedings, 1997
2. Arrundale et al, Handbook and Guide to the Investigation of Clusters of Diseases, Compiled by Leukaemia Research Fund Centre for Clinical Epidemiology, University of Leeds, 1997.
3. Carpenter, Lucy M., Noreen E.S. Maconochie, Eve Roman and D.R.Cox. Examining Associations between Occupation and Routinely Collected Data. J.R.Statist. Soc. A (1997) **160**, Part 3, pp. 507-521.
4. Nicholson, P., Allon, D., McNally, R.J.M. and Rowland, D.J., Analysing the Incidence of Leukaemia in England and Wales, SEUGI 15 Conference Proceedings, 1997.

P.Nicholson
D.Allon

Computing Service
University of Leeds

30th October 1997