Analysing the Incidence of Leukaemia in England and Wales

Paul Nicholson

Introduction

The Leukaemia Research Fund Centre for Clinical Epidemiology at Leeds (LRF), have been engaged in the production of disease atlases of leukaemia in England and Wales since 1990. The current system, based on a DEC Alpha/AXP 2100 running under OpenVMS, employs a variety of programs written in C, FORTRAN, PASCAL and GENSTAT for analysis and ARC/INFO for mapping and is oriented to a purely batch style of working. Communication between the different programs is via intermediate results stored in external files.

The availability of the SAS® system at the University was seen by the LRF as providing a means of integrating the operations involved in the production of the Leukaemia Disease Atlases under the control of a single system. Accordingly, a development project was started, with funding from SAS Institute, with the aim of transferring the Atlas production to the SAS system. In addition, the availability of SAS/GIS (a geographical information system) and the interactive environment of SAS in general, was seen as having the potential of opening up interactive ways of working, more relevant to exploratory work than the batch style of working.

Data

A variety of data are available including case data, population data, coverage information (indicating the time periods during which regions participated in the study and details of any sub-regions not participating), data on diseases and spatial data.

Case data recording the incidence of diseases is referred to the LRF Centre in Leeds from centres distributed widely throughout eight large geographical regions. These regions were carefully selected as the basis for the study in order to achieve sufficient data and to give a balanced coverage of factors considered important in the study of the disease. Items of particular interest are date of birth, address, postcode, diagnosis and disease site.

Yearly population data are taken from the national census data obtained from the Office of National Statistics (ONS). Data are available by sex and age group at various geographic levels such as. county, district, ward and enumeration district.

For a variety of reasons, not all of the areas within the study are included for the full duration of the study. Also, for some regions, only a part of that region (for example, a subset of districts within the county) may have participated in the study. Information on the duration of participation over time and on the geographical completeness of coverage is available for all participating regions. This problem of 'Part Areas' complicates the process of data modelling.

Data on diseases and other medical data is compiled centrally by the LRF in consultation with disease specialists. Diagnoses are often grouped into categories for analysis. Thus, information detailing hierarchies of disease groupings is provided.

Finally, the 1991 Census digitised boundary data sets for England and Wales at county and district level were obtained under licence from the ESRC, and supplied in uncompressed ARC/INFO transport format (E00) for input to SAS/GIS by UKBORDERS located at Edinburgh University.

Application Development

Central to the application is a data model based upon relational theory but expanded to include summary tables at all geographical levels for all area related entities. In addition to improving efficiency in data access, the model addresses problems inherent in the data such as the 'part areas' problem described above.

A graphical user interface is being developed using object-oriented technology provided by the SAS system.
The primary analysis menu, is displayed in Figure 1. The three major boxes correspond to the three

major activities of the user - computation of rates, graphics and statistical analysis.

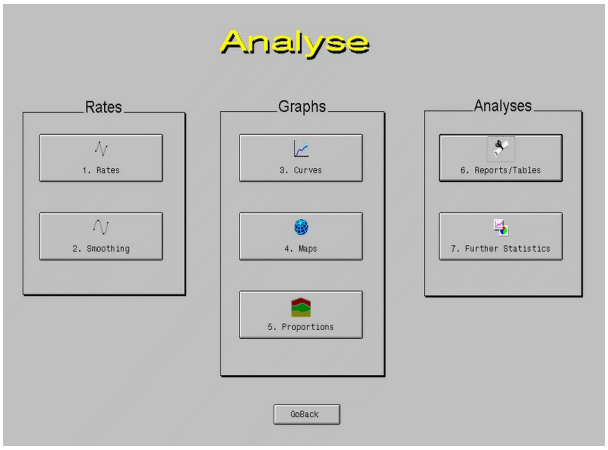

Figure 1. Analysis Menu

The first step for many analyses is the production of tables of age-specific incidence rates. The incidence rate for a particular disease/age-group combination is the ratio of the number of disease occurrences in the age-group to the total number of *person-years* accumulated by the age-group during a specified time period.

To produce such tables, the user must specify a variety of factors. For example, the user may need to define the level of analysis as 'county' , a disease type of NHL, 5 year age groups from 0-79 and start and end dates from within the study time period.

After specifying these criteria, the required subset of data is extracted from the relevant tables and a table of incidence rates is produced.

The Tab Layout object is used to facilitate data selection. The use of a Tab Layout allows the variety of screens required for data selection to be organised within a single frame.

In practice, it is common for multiple studies to be carried out simultaneously, each of which may focus on more than one popluation. The first two tabs provide the means of specifying the study of interest and the particular body of case data required for analysis.

Figure 2 shows the screen used for 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.

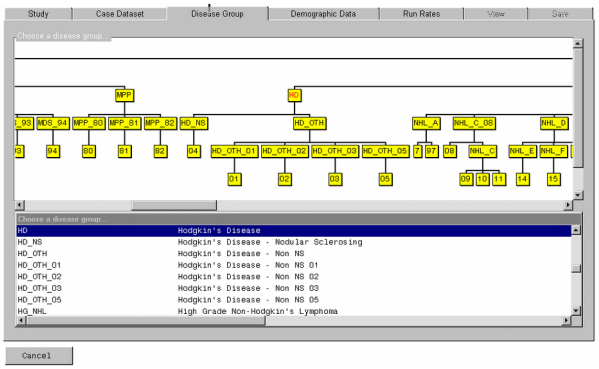

Figure 2. Disease Group Selection Screen

Figure 3 shows the screen used to specify the demographic data selection criteria.

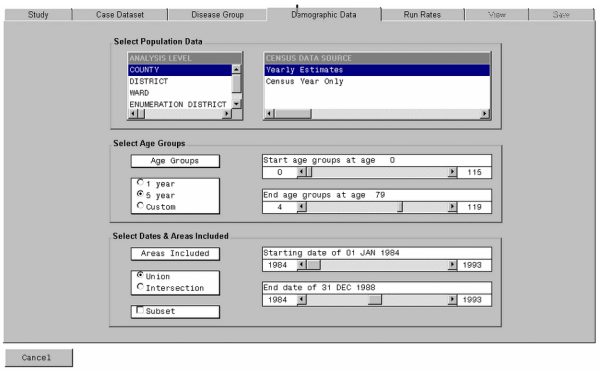

Figure 3. Demographic Data Selection Screen

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. Once the choices are accepted, the application will progress to compute the tables of age-specific morbidity rates.

Analysis

Cancer incidence rates show marked variation with age, and to a lesser extent sex. It is necessary, therefore, to standardise the rates in order to ensure that regional differences are not merely a reflection of differences in the age-sex structure of the population.

To facilitate comparisons between different studies, a direct standardised incidence rate is computed by applying the observed rates to a

reference population whose age and sex distribution is fixed. For the purpose of comparing incidence rates between different areas within this particular study, indirect standardisation is used. Indirect standardisation produces a so-called *Standardised Morbidity Ratio* (SMR), which is expressed as (O/E) x 100 , where O is the observed number of cases and E is the expected number of cases obtained by the standardisation process. Poisson regression is used to compare SMRs between different geographical regions and between different age-groups.

Mapping

It is known that the use of SMRs to map the incidence of a disease can misrepresent the geographical distribution of the incidence of the disease. This can happen when wide variation exists in population sizes between regions. Apparently significantly high SMRs may in fact be based on only a few cases and give a biased picture of disease incidence.

An alternative approach favoured by a number of authors uses empirical Bayes estimation. This approach rests upon hypothesising a prior distribution for the SMR, coupled with an assumption that the observed number of cases follows a Poisson distribution, and exploits Bayes theorem to compute a posteriori estimates of the individual SMRs. Values based upon small numbers of cases are close to the overall mean relative risk whereas values based upon large numbers of cases remain close to the original SMR. The estimates obtained provide a compromise between individual SMRs and the overall mean relative risk.

Figure 4, below, displays the raw SMRs before smoothing has been applied. Figure 5 displays the effect of smoothing based upon a gamma prior distribution for the SMRs.

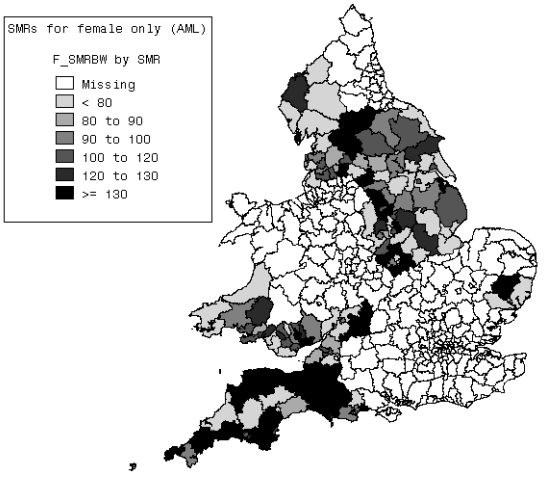

Figure 4. SMRs for Acute Myeloid Leukaemia at district level

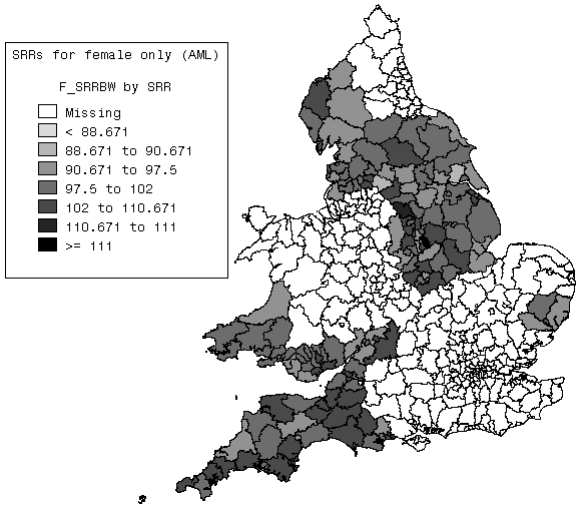

Figure 5. Smoothed SMR estimates obtained using empirical Bayes estimation

The model described above assumes independence in the observed SMRs. However, the smaller the area under study the fewer the number of cases are to be expected. It then becomes more meaningful to take into account possible correlations between SMRs in neighbouring areas. An alternative model, currently under development, allows for a varying degree of spatial autocorrelation by taking into account the SMRs of surrounding areas when estimating the SMR for any area.

Future Plans

Further developments will include the addition of a facility to allow analyses to be deferred for batch processing and the development of facilities for

data administration. In the area of analysis, enhancements planned include the incorporation of further smoothing techniques, spatial statistical tests and seasonal models for SMRs. Further use of SAS/GIS is also planned. In particular, the ability to work with geographic features such as rivers, roads and power lines will enable exploratory analyses of selected clusters to be performed.