DATA MODELLING FOR AN EPIDEMIOLOGICAL DATABASE

AUTHOR: Deborah Allon, University of Leeds
CO-AUTHOR: Paul Nicholson, University of Leeds

ABSTRACT

The dictionary defines Epidemiology as the ‘Science of Epidemics’. In practice, epidemiology is concerned with studying the incidence and transmission of diseases in general, not simply those occurring in epidemics. In epidemiological studies, a variety of types of data can occur. These include case data (giving individual patient details), population data (typically arising from national census surveys), spatial data (e.g. geographic map data sets) and information about diseases. The choice of appropriate data models can be crucial to the successful and efficient access of information required for analysis.

This poster describes the construction of data models for data arising from an epidemiological study of leukaemia. In addition to the types of data described above, additional temporal data are available indicating the periods for which both geographic regions and individuals within those regions were observed during the lifetime of the study. The poster describes some of the problems posed by this complex collection of data, discusses the merits of alternative models and describes the data model ultimately adopted - one based upon a relational model with additional summary tables for improved performance. The majority of analyses carried out require, as input, tables of incidence rates for a particular disease or disease group. The use of PROC SQL and PROC SUMMARY in the construction of these tables is described and examples of the resulting tables, together with graphical representations of the results, are presented.

INTRODUCTION

The Leukaemia Research Fund Centre for Clinical Epidemiology at the University of Leeds (LRF) has been engaged in the production of disease atlases of leukaemia in England and Wales since 1990. Data for the atlases are currently stored in an ORACLE/RDB database on a DEC Alpha/AXP 2100 with analysis and mapping applications carried out using a variety of software systems. The scope of the SAS® system, together with the emergence of SAS/GIS®, was seen as having the potential to provide an alternative solution to the requirements of the LRF - one which would integrate operations under the control of a single system.

Preliminary discussions with the LRF established a number of needs not readily catered for by the existing system. Some of these needs were not included in the list of requirements which the database was originally intended to meet and have arisen as a result of problems affecting the data supply. A thorough re-examination of appropriate data models was seen therefore as an essential pre-requisite to the migration of the application to the SAS system.

To understand how models were designed for the data, it is first necessary to understand the stages in producing the disease atlases and the nature and volume of data required for this epidemiological study area.

ACTIVITY ANALYSIS

The tasks involved in producing a disease atlas can be broken down into three main sections:
Rates

A usual first requirement when analysing incidence data is to calculate tables of incidence rates for a particular disease or group of diseases. For example, the user may specify: a disease group to study (e.g. Acute Myeloid Leukaemia); start and end ages and age groups (e.g. ages 0 to 79 using 5 year age groups); a population data source (e.g. yearly estimates); the level at which results are required (e.g. county); and a time span (e.g. for the first 5 years from a study period of 10 years). Age-specific incidence rates can then be calculated. 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. Incidence rates for leukaemias are very low and it is customary to express them as a rate per 100,000 of population per year, by multiplying the raw incidence rate by 100,000. Figure 1 illustrates a typical table of incidence rates produced for a particular disease or disease group.

<table>
<thead>
<tr>
<th>Area</th>
<th>Age group</th>
<th>No. of female cases</th>
<th>Female rate</th>
<th>No. of male cases</th>
<th>Male rate</th>
<th>No. of total cases</th>
<th>Total rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>05000000</td>
<td>00-04</td>
<td>4</td>
<td>0.98</td>
<td>4</td>
<td>0.87</td>
<td>8</td>
<td>0.93</td>
</tr>
<tr>
<td>05000000</td>
<td>05-09</td>
<td>2</td>
<td>0.50</td>
<td>1</td>
<td>0.29</td>
<td>3</td>
<td>0.39</td>
</tr>
<tr>
<td>05000000</td>
<td>10-14</td>
<td>3</td>
<td>0.57</td>
<td>3</td>
<td>0.72</td>
<td>6</td>
<td>0.64</td>
</tr>
<tr>
<td>05000000</td>
<td>15-19</td>
<td>4</td>
<td>0.76</td>
<td>3</td>
<td>0.61</td>
<td>7</td>
<td>0.69</td>
</tr>
<tr>
<td>05000000</td>
<td>20-24</td>
<td>4</td>
<td>0.71</td>
<td>5</td>
<td>0.92</td>
<td>9</td>
<td>0.81</td>
</tr>
<tr>
<td>05000000</td>
<td>25-29</td>
<td>5</td>
<td>0.96</td>
<td>3</td>
<td>0.53</td>
<td>8</td>
<td>0.75</td>
</tr>
<tr>
<td>05000000</td>
<td>30-34</td>
<td>5</td>
<td>0.96</td>
<td>5</td>
<td>0.98</td>
<td>10</td>
<td>0.97</td>
</tr>
<tr>
<td>05000000</td>
<td>35-39</td>
<td>5</td>
<td>1.05</td>
<td>6</td>
<td>1.14</td>
<td>11</td>
<td>1.10</td>
</tr>
<tr>
<td>05000000</td>
<td>40-44</td>
<td>7</td>
<td>1.39</td>
<td>7</td>
<td>1.51</td>
<td>14</td>
<td>1.45</td>
</tr>
<tr>
<td>05000000</td>
<td>45-49</td>
<td>7</td>
<td>1.61</td>
<td>7</td>
<td>1.75</td>
<td>14</td>
<td>1.68</td>
</tr>
<tr>
<td>05000000</td>
<td>50-54</td>
<td>10</td>
<td>2.75</td>
<td>10</td>
<td>2.69</td>
<td>20</td>
<td>2.72</td>
</tr>
<tr>
<td>05000000</td>
<td>55-59</td>
<td>15</td>
<td>4.17</td>
<td>11</td>
<td>2.85</td>
<td>26</td>
<td>3.50</td>
</tr>
<tr>
<td>05000000</td>
<td>60-64</td>
<td>24</td>
<td>6.50</td>
<td>16</td>
<td>3.93</td>
<td>40</td>
<td>5.16</td>
</tr>
<tr>
<td>05000000</td>
<td>65-69</td>
<td>28</td>
<td>8.48</td>
<td>21</td>
<td>5.49</td>
<td>49</td>
<td>6.86</td>
</tr>
<tr>
<td>05000000</td>
<td>70-74</td>
<td>31</td>
<td>11.70</td>
<td>26</td>
<td>7.52</td>
<td>57</td>
<td>9.31</td>
</tr>
<tr>
<td>05000000</td>
<td>75-59</td>
<td>27</td>
<td>14.57</td>
<td>24</td>
<td>7.86</td>
<td>51</td>
<td>10.45</td>
</tr>
</tbody>
</table>

Figure 1. A typical table of incidence rates for a particular disease group

Analysis

Following the calculation of incidence rates for a disease or disease group, rates are standardised to facilitate comparisons between geographical areas and between age groups in a study, and between studies. A variety of methods are used to adjust the observed rates depending on the analysis performed.

Mapping

The LRF produce choropleth maps, at the county and district levels, depicting incidence of disease in the form of standardised morbidity ratios (SMRs) and smoothed relative risks (SRRs).

DATA ANALYSIS

It was necessary to determine what data is needed to support the LRF’s activities. To simplify understanding, the data can be logically split into three categories:

(i) **Static data**: data that can be used by any study and need only be loaded once. Consists of: area codes (in the form of Small Area Statistics codes (sascodes)) and names; population figures; population standards; geographic map datasets of area boundaries.

(ii) **Study data**: data relating to a particular study and need to be loaded per study. Consists of: study name; details of areas included in the study and their time periods; adjusted populations for any ‘part study areas’; case datasets recording the type of leukaemia as well as the date and location at diagnosis.

(iii) **Application-related data**: data relating to leukaemia and/or the application that could be used by more than one study. Consists of: disease codes, names and groupings; biopsy codes, names and groupings; commonly used region definitions (area groupings).

Studying the data revealed some important characteristics warranting further consideration:

- ‘Part Areas’
- Study areas changing over time
- Changes to regional boundaries during the study period
- The way in which population data is made available

A ‘part area’ is an area of the country for which case data was not collected for one or more of its components during the collection period. (e.g. Avon was a part county in a study because one of its districts was omitted). It is important to emphasise that the ‘part area’ syndrome is governed by the study and not the areas of the country.

Areas and part areas included in a study could change over time during the course of a study. For example, the county named ‘Avon’ could be a ‘part area’ for the first few years of a study, due to one of its districts not participating in the study, but could be considered as a ‘whole area’, at a subsequent time, when the absent district enters the study.

Over the period of a study, changes to regional boundaries may take place requiring changes to be made to the Small Area Statistics codes issued with the census data describing the various regions. Assumptions may therefore be required about the boundary data that is to be used.

Census data is released, by the Office of National Statistics (ONS), at the highest regional level first (i.e. county, followed by district, ward and enumeration district level). It can take considerable time after a census to release full ward and ED level population figures.

MODELS CONSIDERED

I - Demographic Data

Initially three data models were proposed, based on relational database theory.

The first model considered was a direct physical implementation of the logical data model, with one table for all the ‘area’ information, one table for all the ‘study area’ data, and one table for all the ‘area age group populations’. Figure 2 displays the data model.
Model 1: Advantages

- Has flexible file structures, especially for area data, and could handle any form of area data e.g. health authority regions, or those based on postcode.
- The hierarchy of areas is reflected in information in the ‘area’ table, and does not need to be coded in separately.
- Both ‘study area’ and ‘area’ entities would be a collection of data summarised at different levels. This could also be stored as separate tables for speed of access.
- Would be flexible and fast but large.
- A ‘part area’ field for ‘study area’ would enable both the aggregation from lower-level data where necessary and the use of higher-level summary data when available for efficiency.

Model 1: Disadvantages

- Large tables could present a performance issue.
- There are no fields for storage of details relating only to wards. If the file structure for areas is amended to include these, then these fields would be null for districts and counties for example, so wasting storage space.

The second model considered differs from Model 1, in that the ‘area’ information is separated into different tables, one for each level. ‘Area age group populations’ are only held at the lowest level (ward), as populations at higher levels can be calculated from these. This model allows details which are important at ward level to be held in the appropriate place, with no redundancy. Figure 3 shows the data model.
Model 2: Advantages

- Only the lowest level of populations are stored (as in true relational form). Summaries at other levels are aggregated from the lowest level, the advantage being that we have the same processing for all areas whether the area is ‘part’ or not.
- It is not necessary to hard code the names of the regions into the file structure (such as district code or county code). Level numbers could be used instead, ensuring that the file structure remains flexible enough to handle other types of area data (e.g. health authority areas).
- Can store ‘ward only’ attributes in the correct place.

Model 2: Disadvantages

- Would be slow without ready summary tables, as there is a lot of aggregating to be done unless working at ward level.
- Hard coding of district and county field names into the file structures creates an inflexible system as regards handling other area data.
- An assumption needs to be made about what is deemed the lowest level of ‘area’. Further problems would be encountered if these areas were ever ‘part’.

The third model considered is almost identical to Model 2 except that the composite sascode is replaced by a separate key for each stratum. This permits any coding scheme to be adopted for area information and does not restrict the user to the use of sascodes.

II - Disease Group Data

The data relating to disease and disease groups presented a different set of problems. The main requirement was to be able to identify both the membership of a disease group, in order to produce appropriate subsets of cases relating to a disease group for subsequent analysis, and to identify the group(s) to which a particular disease belongs.

Initially, the relational model was investigated to see if it could provide a solution to the storage of disease codes and group information.

In this application concerning leukaemia, a disease group contains many diseases, and a disease can belong to many disease groups. This can be represented by a many-to-many relationship as depicted in Figure 4.

![Figure 4. Many-to-many relationship](image)

Many-to-many relationships often conceal other entities not yet identified and indeed this many-to-many relationship can be dissolved by the use of an intersecting entity to reveal three entities about which attributes can be kept, as shown in Figure 5.

![Figure 5. One-to-many relationships](image)
Example attributes and attribute values of relations such as these are displayed in Figure 6. (Field names are in bold and key fields are underlined).

**Disease**

<table>
<thead>
<tr>
<th>Disease Code</th>
<th>Disease Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>Acute myeloid leukaemia</td>
</tr>
<tr>
<td>56</td>
<td>AML (M1) without maturation</td>
</tr>
</tbody>
</table>

**Group Member**

<table>
<thead>
<tr>
<th>Disease Code</th>
<th>Group Belonged To</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>AML</td>
</tr>
<tr>
<td>56</td>
<td>AML_M1</td>
</tr>
<tr>
<td>56</td>
<td>AML_M1_M3</td>
</tr>
<tr>
<td>56</td>
<td>AML_M1_M7</td>
</tr>
</tbody>
</table>

**Disease Group**

<table>
<thead>
<tr>
<th>Group Name</th>
<th>Long Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML</td>
<td>Acute Myeloid Leukaemia group</td>
</tr>
<tr>
<td>AML_M1</td>
<td>Acute Myeloid Leukaemia subgroup M1</td>
</tr>
<tr>
<td>AML_M1_M3</td>
<td>Acute Myeloid Leukaemia subgroups M1_M3</td>
</tr>
<tr>
<td>AML_M1_M7</td>
<td>Acute Myeloid Leukaemia subgroups M1_M7</td>
</tr>
</tbody>
</table>

Figure 6. Example table contents for disease-group information

These relations provide the necessary names of codes and groups that may be needed for reports etc., and also show the membership of all the disease groups.

Simple SQL statements facilitate the identification of individual disease codes belonging to a group or conversely the identification of groups to which a particular disease belongs.

E.g.1. To retrieve all members of the group ‘AML’, we would specify :-

\[
\text{select discode from work.grpmem where group='AML'};
\]

E.g.2. To identify the groups to which disease code 56 belongs, we would specify :-

\[
\text{select group from work.grpmem where discode='56'};
\]

Therefore, these relations satisfy one of our two criteria - that of subsetting case data for a particular disease. However, these tables do not reflect the hierarchy of the disease groups and do not present a graphical means of presenting groups to the user other than presenting an ordered, but flat, list of group names.

Another table can be introduced specifically to represent the hierarchy which is not apparent from the relations above. This table would possess a structure such as that shown in Figure 7, where information about the parent of every disease group or code is provided.
### Figure 7. Disease group hierarchy

Another way of storing hierarchical disease group information would be to use lists. Nested sublists could be used to populate nested listboxes or popmenus, to provide a graphical means of selecting a disease group. However, the SCL code needed to manipulate the complex lists needed for this disease area would be tortuous, although not impossible, to write.

A more suitable mechanism for representing hierarchical data, and of providing the user with a graphical front end for observing and choosing disease groups, is the new Organisational Chart object in SAS/AF® released in SAS 6.11.

A dataset such as that displayed in Figure 7 can be used to populate an organisational chart, such as that pictured in Figure 8.

### Figure 8. Organisational chart displaying a disease hierarchy

An organisational chart provides the user with a graphical representation of the disease hierarchy and also provides a mechanism for data selection. Once the user has selected a group, the disease codes that constitute that group can be identified by fetching all the leaf nodes beneath the group node.

It is likely that the methods associated with the organisational chart class will provide the most efficient means of determining the disease codes belonging to a selected disease group, since the organisational chart object is held in memory. However, in other parts of the application, where the need to determine group membership may arise, the use of SQL to access this information from data sets may prove to be more efficient.
It is not uncommon for the disease hierarchy to require modifications. In this event, the organisational chart has the potential to provide a powerful tool for visual editing of such hierarchical data. Early experiments with the existing methods provided with the organisational chart indicate that the dynamic modification of a disease hierarchy is feasible, although the provision of further methods would simplify the SCL programming required.

SOLUTIONS

I - Demographic Data

In a normalised relational data model summary information which can be calculated is not stored. For example, county level populations can be calculated from ward level populations, and time periods for counties can be calculated from ward time periods, purely by storing the minimum of information at ward level and recording which wards are in which districts and counties. However, this ‘bottom up’ approach has serious shortcomings in this application area.

In reality, if a new system is to be satisfactory to the user, the user must be able to access the data required with minimum effort and minimum delay. If a particular study performed analyses at county level only and there were no ‘part areas’, it would be unnecessary to store data at any level other than county. It may not even be possible to store data at lower levels in some studies, as data at these levels may not be available, nor ever be required. If data is available at different levels then it is still possible to aggregate up from the lowest levels where desired.

Analyses are generally performed at the highest geographical level first (e.g. county) and progress down through the hierarchy. Also, population figures are released at the highest level first and it can take considerable time after a census is completed to release full ward and ED level population figures. Thus, a data model that could only make use of the lowest level data would be unusable until quite a while after the completion of a census.

It was evident therefore that a data model was required that would allow data to be stored and used at any level.

Figure 9, below, provides a pictorial representation of the data model, around which system development began, implemented as tables in SAS. The model was based on two of the earlier proposed models (derived using relational theory, entity modelling and normalisation) and expanded to include relevant summary tables at all geographical levels for all area-related entities (‘area’, ‘study area’, ‘area age-group population’ and ‘adjusted study area age group population’). The solid horizontal line in Figure 9 represents the dividing line between the normalised part of the model and the summary tables. The summary tables above the line could all be calculated from the tables below the line if these were complete. Although this model does not support a truly flexible file structure for area data, such as Model 1 did, it was felt to be the best starting point to address the needs of the LRF and the problems mentioned earlier.

As well as being necessary, storing area related information at all geographical levels addresses the problems of ‘part areas’ and ‘study areas changing over time’. If we know that a county was ‘part’, we can drill down to the geographical level below (in this case district) to find out which districts were included and for which dates. If any of these were part, we can drill down again, and so on. Therefore, there can be complete information in the ‘study area’ tables for all the areas participating in a study and the dates thereof.
Figure 9. Pictorial representation of the proposed data model

The data model described allows SQL (in the form of SUBMIT CONTINUE blocks from Screen Control Language) to be used to extract the data required for the construction of age-specific incidence rates tables. The two main strands in this calculation are the subsetting of cases and the collection of person years for each area and age group. The actual SQL code employed is too lengthy to show in full here but Figure 10, below, illustrates a typical piece of SQL to construct a rates table.

```sql
/*Create rates table*/
create table work.rates as
    select c.sas_code, c.age_gp, (c.fcases/p.fpyrs)*100000 as f_rate,
        c.mcases, (c.mcases/p.mpyrs)*100000 as m_rate,
        (c.fcases+c.mcases) as tcases,
        ((c.fcases+c.mcases)/(p.fpyrs + p.mpyrs))*100000 as t_rate
    from work.fullfm as c, work.pyears as p
    where c.sas_code=p.sas_code and
        c.age_gp=p.age_gp
    order by c.sas_code, c.age_gp;
```

Figure 10. SQL code to create a rates table

A common requirement is to be able to generate a multiplicity of tables in one session (e.g. one table for every area of every stratum). The production of very large volumes of output is at odds with the nature of interactive working. Accordingly, a batch facility is planned to cater for this. It is likely that PROC SUMMARY will be used as an alternative to SQL for this purpose. Using PROC SUMMARY, the full set of tables for every area of every stratum (n-way crossing) can be produced and stored in a single data set. A simple report writing step is all that is required, subsequently, to extract each individual table for presentation.
II - Disease Group Data

For disease group data, both the tables and the organisational chart, as discussed earlier, are employed to facilitate the selection of disease groups required for analysis.

CONCLUSION

Implementing an application without giving consideration to different ways of storing the data may result in inefficiencies in both data storage and data access mechanisms. Studying the data ensures that the different entities and the relationships between them are identified, enabling a sensible data model to be developed. The code to access and extract data from a well-organised data model is likely to be much simpler than the often complicated code required to access badly organised data.

The data model described in this poster was successfully implemented in the application under construction and allowed data to be accessed efficiently using the SQL procedure. The extraction of data is now flexible and any selection criteria can be combined to calculate incidence rates for a subset of data.

This paper illustrates a solution to the design of data models for epidemiological data, to satisfy the requirements of a user dealing with leukaemia data. It is hoped that in the future the model can be adapted, if necessary, to cope with the requirements of other epidemiological studies.

REFERENCES


ACKNOWLEDGEMENTS

We would like to thank SAS Institute for providing both financial and technical support for this project and the Leukaemia Research Fund for providing financial support for the Data Collection Study. This work is based on data provided with the support of the ESRC. Thanks are due to the Office of National Statistics for the provision of population estimates, which remain the property of the Crown, and also to the staff of UKBORDERS, Edinburgh, for the supply of the 1991 Census digitised map boundary datasets for England and Wales.

SAS, SAS/AF and SAS/GIS are registered trademarks of SAS Institute in the USA and other countries. © Indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.
ADDRESS FOR CORRESPONDENCE

The authors may be contacted at the following addresses:

Post: University Computing Service
      University of Leeds
      Leeds
      LS2 9JT
      UK

Email: D.Allon@leeds.ac.uk
       P.Nicholson@leeds.ac.uk

Tel:   +44 113 233 5414 (D. Allon)
      +44 113 233 5405 (P.Nicholson)

Fax:   +44 113 233 5411

A project web page is located at:
http://www.leeds.ac.uk/ucs/projects/sas/epi.html