A SAS-AUTOGRP INTERFACE FOR THE ANALYSIS OF HOSPITAL DISCHARGE DATA

Daniel H. Freeman, Jr., Michael B. Bracken, Robert F. Elia, and Jean L. Freeman, Yale University

1. Introduction

Since the middle of the 1960's a variety of health related data sets have become available to the health analyst. Of particular importance is that these data sets are in machine readable form so that extensive computer assisted analysis is feasible. In this paper a typology of these data sets is presented along with a brief summary of some of their peculiar difficulties. Then a system designed for this type of data is discussed. Finally, an example, which illustrates a strategy for analysis is presented. It uses data from the Hospital Discharge Survey (HDS) of the National Center for Health Statistics (NCHS).

Traditionally health data have been gathered either as a by-product of administrative record keeping systems or as part of a well specified scientific study. The administrative approach has been quite successfully employed in such diverse areas as the Vital Registration System which is based on the systematic certification of vital events and in non-profit censuses of hospital records done by the Professional Activity Survey (PAS). Vital registration system data have been widely utilized for establishing basic health parameters in the general population. The PAS data set has been used primarily for establishing norms in hospital utilization.

An alternate means of data collection is a scientifically designed specific hypothesis. There are three types of scientific study used in health research. The most frequently conducted is the case-control study where subjects are stratified according to the presence or absence of a disease of interest. Data are gathered on the subjects and analyzed to permit the identification of risk factors associated with the disease. This type of study parallels the cohort study. In a cohort study, people with a common life event, such as date of birth, are followed over time. The cohort is then post-stratified according to some exposure factor and the strata are compared according to the levels of the identified events. Again, the relationship between a risk factor and the disease can be estimated. The third data collection design is the clinical trial. Here, individuals are randomly allocated to various treatment regimens and the resulting effects can be compared using standard statistical methods. A common feature of these methods of gathering data for scientific studies is that the objectives are usually narrowly defined, and alternative outcomes can be specified in advance. Similarly the administrative data system is justified in terms of improving efficiency, economy, or other internal criterion. As a result subsequent use of the data for academic purposes is comparatively unimportant.

An alternative data system which has become quite wide-spread is the "multi-purpose survey". This approach gathers data from a population through the use of probability sampling methods. The data gathering process is intended to provide information for a variety of users. The best known of these, in the health field, is the National Health Survey (NHS). The NHS actually includes a variety of sample surveys each of which have two important characteristics. First they provide data about a carefully delineated cross-section of specified target populations. Second, since they provide data to a variety of users their underlying rationale is sometimes less apparent than that of administrative or scientific data systems. However, while these surveys are of great potential value, their general and scope lead to a number of difficulties.

The most immediate problem is the sheer magnitude of the health survey. Typically they result in large numbers of records (length) and possibly hundreds of variables (width). For example the Hospital Discharge Survey annually generates more than 200,000 records; the Health Interview Survey generates up to 130,000 records, depending on the aspect understudy; the Health and Nutrition Survey contains data on over 30,000 individuals requiring 20 tapes for storage. Needless to say, data sets of this size cannot be casually examined at reasonable cost.

This problem of size is intimately related to the difficulty of accessing these data. It is self-evident that such data sets can only be dealt with through the use of powerful data processing equipment. Unfortunately, the availability of a data tape and an appropriate machine is not a sufficient condition for an analysis to be feasible. An important part of substantive and statistical knowledge to integrate the results of various tabulations and build on this to generate progressively more informative tabulations. In this age of specialization and high technology the analyst is usually not an expert in computer science and thus must rely either on programmers or general purpose statistical packages.

A third problem is that while specific hypotheses may be tested with these data sets, the state of knowledge about health is such that this may be a comparatively artificial exercise. For example, even if the etiology of a particular disease or other morbid condition is well known the population characteristics may remain a mystery. As a specific example, which will be considered further, acute spinal cord lesion is almost always due to external trauma. But most estimates of its incidence and associated factors are based on a single study in Northern California. Thus, it may be of interest to build an empirical model with the available national survey data.
In addition to problems associated with the size and accessibility of the National Health Survey there are problems unique to each of its component surveys. For example, if one is to use the HDS then there are important limitations with respect to methods of data collection, specificity of data, and validity of resulting statistical analysis. The HDS method of data collection is a probability sample, thus it must be recognized that there are sampling weights and errors which should be incorporated in the analysis. Discussion of methods for doing this are found in NCHS (1970) and Koch, et al. (1975). Moreover the population of hospitals covered changes continuously so that the sampling frame has required repeated updating. Finally, despite significant efforts at quality control, the data may be subject to an overall 10 percent level of coding errors (NCHS 1970: 13-15).

Limitations in specificity and validity also apply to the HDS. Specificity can mean either that diagnostic codes are too general or that general hospital practice is not to use all the available ICD-A refinements. Similarly, the use of medical abstracts ultimately relies on the validity of information contained in the medical record. Since medical records are prepared by thousands of individuals, it is inevitable that misreporting and ambiguous interpretations will occur. Thus one must utilize and interpret data from health surveys with a great deal of caution.

Faced with this host of difficulties and limitations there is a great temptation to simply discard the data and find more direct approaches to the analysis of health problems. However, the potential benefits of the multipurpose survey must be considered. Specifically, the data are available and can be accessed at low cost, if one uses an appropriate analysis strategy. For example the purchase cost to the user works out to about $0.001 per record. Second, the various surveys cover long and interesting periods of time. Two obvious examples are HDS and the Health Interview Survey. Each provides a permanent record of health patterns in the United States since the inception of Medicare and Medicaid. A third benefit is that these surveys permit inferences to well defined populations, thus strengthening the theoretical base for conclusions based on them. These populations may be individuals or patient records, but for NMS they are always national in scope.

2.0 A Package for Accessing Large Data Bases

The AUTOGRP package is actually an extension of a general purpose programming system called CML (Conversational Modelling Language) (Mills et al. 1976). CML is an interactive algorithmic language similar to PDL, which statements may be executed interpretively or compiled into programs. The linguistic capabilities of CML include a meta-language for the specification of new or modified command structures. This mechanism permits the user to define unique sets of commands. This has the effect of lowering the cost of using the CML system to his particular needs. Moreover, when using an existing CML program or package, the entire system remains at his disposal. Thus, he may invoke both his defined command set and the standard CML command set.

If the user's dataset does not meet the database structure requirements of AUTOGRP, it may be reorganized in a two stage process applying two other CML extensions - VCS (the Variable Conversion System) and AUTOSEL (Theriault et al. 1979: chapter 6; 1). VCS provides a means by which a database containing card image data may be used to produce a "converted" database, with all numeric variables transformed to either integer or floating point binary representation.

Once the entire database has been converted with VCS, the user invokes AUTOSEL to create a direct access disk subfile containing only the subset of records and variables required for the analysis. This is a two step process. First, the user interactively enters the necessary information that specifies the subfile, and appropriate job control is constructed. Second, the job control is submitted for batch processing and a file formed according to the specifications in Step 1. These subfiles may then be analyzed with either AUTOGRP or SBS (Ella and Freeman 1975).


The foregoing discussion can be illustrated with an example taken from HDS. It will be used since it is readily available and parallels in structure many administrative data sets. It is large and varied enough to illustrate some of
the difficulties which occur in the analysis of health data. HDS is also the most appropriate of the National Health Surveys for studying the incidence of rare but generally non-fatal diseases. The specific analysis will be based on the HDS data sets for 1971 through 1975. Taken together this contains in excess of 1,000,000 records randomly selected from hospitals across the country.

The method of sample selection is well documented by NCNS (1970) and the selection design is summarized in the annual reports (eg. NCHS 1975). However, the highlights of the current design should be reviewed before an analysis is undertaken:

"The information for the survey is abstracted from the face sheets of the medical records sampled for inpatients discharged from a national sample of the non-Federal general and special short-stay hospitals... In 1975 there were approximately 232,000 medical records sampled from 432 hospitals that participated in the survey (NCNS 1975)".

The target population includes all hospitals of 6 or more beds and an average patient length of stay of under 30 days. The selection plan is stratified according to the 4 census regions and 7 hospital bed sizes. Within each stratum controlled selection is done so as to insure representation of the various ownership types. Thus HDS may be used to produce statistically valid estimates for the population of hospital discharges on an annual basis. The steps in breaking down a data set such as this is are illustrated in Figure 1.

An almost endless variety of analyses are feasible for a data set such as this. We chose to use it to study the factors associated with the incidence of acute spinal cord injury (SCI). The first step was to identify the diagnostic codes which were likely to yield all or most of the discharges with this disorder. The work of Kraus et al. (1975) and Webb et al. (1978) suggests that the ICDA-8 codes 344, 805, 806, 958 were appropriate. Using this for the first pass the AUTOSEL program was used to form the appropriate subset of the data.

Data from the National Hospital Discharge Survey were converted with VCS and stored in five separate tape datasets corresponding to the years 1971 through 1975. The AUTOSEL utility was used to create subfiles from each year with data necessary for the analysis, namely all records with a first, second, third, fourth, or fifth listed diagnosis in the ICDA8 code ranges 3440-3449, 8050-8069, 9580-9589. A CM program was written to merge the five subfiles into a single sequential dataset. A CM utility program was then used to rewrite the sequential file into a direct access file suitable for use by either AUTOGRP or SAS. The resulting card image data set contained 1,104 records. At this point descriptive statistics were generated using the AUTOGRP package. The most important of these were histograms for length of stay and the distributions by listed diagnosis. The latter is shown in Figure 2.

The most obvious conclusion one obtains from Figure 2 is that the overwhelming majority of the discharges are for codes of 344 and 805. Webb et al. (1978) noted that these codes contained fewer than 4% of the SCI patients. A second factor which must be considered is that often the SCI patients are transferred among acute care hospitals. In order to take account of this it was felt that if patients were discharged alive within two weeks of admission they were likely to represent multiple admissions rather than true incidence. (Bracken et al. 1979).
Figure 2. Output from AUTOPR

*/Frequency of spinal cord discharges as listed by primary (ddx1), secondary (ddx2), tertiary (ddx3), fourth (ddx4), and fifth (ddx5)
*/
*/diagnoses. Total discharges in file=11404. 370 discharges have
*/multiple spinal cord listings.
*/
*/#display of primary diagnoses:
ddx1 of # we "ddx1 805,806,344,958"
*/
*/ 4 CELLS 4137 OBSERVATIONS
*/
*/NUMBER      MEAN   S.D  DDX1
*/OBS.   PCT  LOS  LOS  CODE
*/ 980  23.69%  24.12  36.79  344 OTHER CEREBRAL PARALYSIS
*/2627 63.55%  12.34  12.21  805 FRACTURE AND FRACTURE DISLOCATION OF VERTEBRAL COLUMN WD
*/103  2.64%  42.31  54.29  804 FRACTURE AND FRACTURE DISLOCATION OF VERTEBRAL COLUMN WI
*/425 10.27%  15.36  28.22  958 SPINAL CORD LESION WITHOUT EVIDENCE OF SPINAL BONE INJUR
*/
*/#display of secondary diagnoses:
ddx2 of # we "ddx2 805,806,344,958"
*/
*/ 4 CELLS 4459 OBSERVATIONS
*/
*/NUMBER      MEAN   S.D  DDX2
*/OBS.   PCT  LOS  LOS  CODE
*/ 3905 90.40%  20.73  27.19  344 OTHER CEREBRAL PARALYSIS
*/577 13.39%  13.66  20.12  805 FRACTURE AND FRACTURE DISLOCATION OF VERTEBRAL COLUMN WD
*/31  0.70%  22.45  65.27  804 FRACTURE AND FRACTURE DISLOCATION OF VERTEBRAL COLUMN WI
*/246  5.52%  16.42  27.27  958 SPINAL CORD LESION WITHOUT EVIDENCE OF SPINAL BONE INJUR
*/
*/#display of tertiary diagnoses:
ddx3 of # we "ddx3 805,806,344,958"
*/
*/ 4 CELLS 1779 OBSERVATIONS
*/
*/NUMBER      MEAN   S.D  DDX3
*/OBS.   PCT  LOS  LOS  CODE
*/ 1403 78.86%  18.21  21.34  344 OTHER CEREBRAL PARALYSIS
*/259 14.56%  14.31  13.79  805 FRACTURE AND FRACTURE DISLOCATION OF VERTEBRAL COLUMN WD
*/87  1.56%  14.70  16.69  806 FRACTURE AND FRACTURE DISLOCATION OF VERTEBRAL COLUMN WI
*/107  6.01%  12.33  15.49  958 SPINAL CORD LESION WITHOUT EVIDENCE OF SPINAL BONE INJUR
*/
*/#display of fourth diagnoses:
ddx4 of # we "ddx4 805,806,344,958"
*/
*/ 4 CELLS 915 OBSERVATIONS
*/
*/NUMBER      MEAN   S.D  DDX4
*/OBS.   PCT  LOS  LOS  CODE
*/ 727 79.45%  18.07  17.60  344 OTHER CEREBRAL PARALYSIS
*/134 14.64%  16.99  17.03  805 FRACTURE AND FRACTURE DISLOCATION OF VERTEBRAL COLUMN WD
*/5  0.55%  13.20  22.94  806 FRACTURE AND FRACTURE DISLOCATION OF VERTEBRAL COLUMN WI
*/49  5.36%  17.67  28.01  958 SPINAL CORD LESION WITHOUT EVIDENCE OF SPINAL BONE INJUR
*/
*/#display of fifth diagnoses:
ddx5 of # we "ddx5 805,806,344,958"
*/
*/ 4 CELLS 604 OBSERVATIONS
*/
*/NUMBER      MEAN   S.D  DDX5
*/OBS.   PCT  LOS  LOS  CODE
*/ 390 80.38%  18.00  19.38  344 OTHER CEREBRAL PARALYSIS
*/81 14.74%  17.27  17.40  805 FRACTURE AND FRACTURE DISLOCATION OF VERTEBRAL COLUMN WD
*/2  0.21%  20.00  0.00  806 FRACTURE AND FRACTURE DISLOCATION OF VERTEBRAL COLUMN WI
*/12  2.40%  9.42  9.67  958 SPINAL CORD LESION WITHOUT EVIDENCE OF SPINAL BONE INJUR
*/
Based on these criteria, data were again subset.

The second subsetting was limited to ICD-8 codes 806 and 958. In addition, if the patients, were discharged alive the length of stay had to be at least fifteen days. This second data reduction could be achieved either with SAS (Barr et al. 1976) or AUTOGRP. It was more convenient to use SAS since a SAS data set could be generated cheaply as a by-product of the SCI incidence study which will be presented in another paper. The SAS data set that resulted contained 319 records of which 16 were persons under age 15 at admission. This completed the first two levels of analysis which may be thought of as problem specification and refinement. The final step is the construction of an empirical model.

It was of particular interest to ascertain those factors which effect the survival to discharge of the SCI patient. The available factors and the corresponding levels are shown in Table 1. Since there were only 16 patients under 15, this group was not considered further. Each variable was cross-tabulated with discharge status and the resulting tests of association are also shown in Table 1. Clearly, race is the factor most strongly associated with patient survival. A variable selection algorithm was then implemented by stratifying the data according to the levels of race. Each of the remaining variables were then cross-classified with outcome within each stratum. The results are shown in Table 2, where age is the variable most strongly associated with outcome after race is controlled. This process is repeated again stratifying on both age and race. The results are shown in Table 3. At this step no variable is associated with outcome at the usual levels of significance so the process is terminated. This process of variable selection to form a multidimensional contingency table is similar to that suggested by Higgins and Koch (1977). The factors age, race, and year are thus selected for forming a 4-dimensional cross-classification against the response, outcome. The analysis of this table can easily be performed using any of the standard approaches to categorical data analysis (e.g., the program 3F in the BMDP package, Dixon and Brown 1977). In this case it turns out that age is inversely related to survival but the effect of race depends on what survey year is considered. This analysis will be discussed in a later publication.

4.0 Conclusion

The limitations and potential benefits of data sets from the National Health Survey especially HDS, have been reviewed in this paper. Before the benefits can be realized, it is apparent that these datasets must be accessible to substantive analysts at reasonable cost. Up until recently this has been a virtually insurmountable problem. During the last few years this has been reduced in two ways. First, high speed computer packages having a high level of flexibility are now widely available. This flexibility includes both the ability for unique user written programs to be linked to packages, such as AUTOGRP and SAS, and compatibility with existing statistical packages such as SAS. The second development is a formal SAS-AUTOGRP interface. This makes it possible to exploit the best features of several different data management and analysis packages. In the present example, AUTOGRP and AUTOSEL were used to create modest size datasets and generate descriptive statistics. The latter were used to refine the problem understudy and produce a more disease specific dataset. Subsequently, the data management and analysis capabilities of SAS became more important for the development of an empirical model. Finally, the relatively advanced statistical capabilities of a package such as BMDP were exploited to finalize the analysis.

Acknowledgements

This study was supported in part by the National Institute Neurological and Communicative Disorders and Stroke Contract Number No1-NS-7-2361. And thank you to Robyn Cirasuolo for her Administrative support.

References


Table 1. Associations with discharge status among ICDA-8 codes 806 and 958. Discharges restricted to LOS > 14 days, if alive and ages over 14. (NCHS-HDS-1971-1975)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Levels</th>
<th>n</th>
<th>df</th>
<th>G^2</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (A)</td>
<td>15-29, 30-44, 45-64, 65+</td>
<td>300</td>
<td>3</td>
<td>10.10</td>
<td>0.02</td>
</tr>
<tr>
<td>Sex (S)</td>
<td>Male, Female</td>
<td>300</td>
<td>1</td>
<td>0.69</td>
<td>0.41</td>
</tr>
<tr>
<td>Race (R)</td>
<td>White, Other</td>
<td>270</td>
<td>1</td>
<td>16.50</td>
<td>0.00</td>
</tr>
<tr>
<td>Marital Status (M)</td>
<td>Married, Single, Other</td>
<td>288</td>
<td>2</td>
<td>9.02</td>
<td>0.15</td>
</tr>
<tr>
<td>Region (R)</td>
<td>East, Central, South, West</td>
<td>300</td>
<td>3</td>
<td>5.56</td>
<td>0.14</td>
</tr>
<tr>
<td>Bedsize (B)</td>
<td>6-299, 300-499, 500+</td>
<td>300</td>
<td>2</td>
<td>0.28</td>
<td>0.87</td>
</tr>
<tr>
<td>Ownership (O)</td>
<td>Government, Non-profit, Other</td>
<td>300</td>
<td>2</td>
<td>0.12</td>
<td>0.94</td>
</tr>
<tr>
<td>Year (Y)</td>
<td>1971, 1972, 1973, 1974, 1975</td>
<td>300</td>
<td>4</td>
<td>3.39</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Notes:
1. n = sample size without missing values
2. df = degrees of freedom
3. G^2 = likelihood ratio chi-square
4. p-value = 1-P(G^2 < g^2) if G^2 follows a chi-square distribution
5. Total possible cells = 2x4x2x2x3x4x3x3x3x3 = 17280
Table 2. Associations with discharge status controlling for race.

<table>
<thead>
<tr>
<th>Level of race</th>
<th>Crossed variable likelihood ratio chi-square ($G^2$)</th>
<th>Age</th>
<th>Sex</th>
<th>Marital status</th>
<th>Region</th>
<th>Bed size</th>
<th>Ownership</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>3.60</td>
<td>0.06</td>
<td>0.66</td>
<td>2.78</td>
<td>4.56</td>
<td>1.85</td>
<td>7.35</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>13.16</td>
<td>0.04</td>
<td>6.23</td>
<td>5.60</td>
<td>4.09</td>
<td>3.28</td>
<td>8.25</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16.76</td>
<td>0.10</td>
<td>6.89</td>
<td>8.38</td>
<td>8.65</td>
<td>5.13</td>
<td>15.60</td>
<td></td>
</tr>
<tr>
<td>df</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>p-Value</td>
<td>0.01</td>
<td>0.96</td>
<td>0.14</td>
<td>0.21</td>
<td>0.07</td>
<td>0.27</td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Associations with discharge status controlling for age, race

<table>
<thead>
<tr>
<th>Level Age of Race</th>
<th>Crossed variable likelihood ratio chi-square ($G^2$)</th>
<th>Sex</th>
<th>Marital Status</th>
<th>Region</th>
<th>Bed size</th>
<th>Ownership</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-29 White</td>
<td>1.30</td>
<td>1.84</td>
<td>3.97</td>
<td>0.31</td>
<td>3.89</td>
<td>6.17</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0.62</td>
<td>2.82</td>
<td>1.52</td>
<td>0.88</td>
<td>1.72</td>
<td>3.22</td>
<td></td>
</tr>
<tr>
<td>30-44 White</td>
<td>0.13</td>
<td>1.43</td>
<td>1.38</td>
<td>3.86</td>
<td>1.49</td>
<td>1.75</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>1.48</td>
<td>1.88</td>
<td>3.84</td>
<td>1.07</td>
<td>5.84</td>
<td>7.66</td>
<td></td>
</tr>
<tr>
<td>45-64 White</td>
<td>2.66</td>
<td>0.58</td>
<td>7.16</td>
<td>5.52</td>
<td>1.33</td>
<td>5.02</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0.00</td>
<td>0.17</td>
<td>4.97</td>
<td>6.17</td>
<td>5.34</td>
<td>4.30</td>
<td></td>
</tr>
<tr>
<td>65+ White</td>
<td>0.13</td>
<td>0.19</td>
<td>2.97</td>
<td>1.60</td>
<td>0.38</td>
<td>9.98</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2.09</td>
<td>4.87</td>
<td>0.37</td>
<td>0.37</td>
<td>0.00</td>
<td>4.87</td>
<td></td>
</tr>
<tr>
<td>Total G^2</td>
<td>8.41</td>
<td>13.78</td>
<td>26.18</td>
<td>19.78</td>
<td>17.99</td>
<td>42.97</td>
<td></td>
</tr>
<tr>
<td>df</td>
<td>8</td>
<td>16</td>
<td>24</td>
<td>16</td>
<td>16</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>p-Value</td>
<td>0.39</td>
<td>0.62</td>
<td>0.34</td>
<td>0.23</td>
<td>0.32</td>
<td>0.09</td>
<td></td>
</tr>
</tbody>
</table>