Applying Data Mining Techniques to a Health Insurance Information System

Marisa S. Viveros
IBM Research Division
T.J.Watson Research Center
Yorktown Heights, NY 10598
viveros@watson.ibm.com

John P. Nearhos
Health Insurance Commission
Tuggeranong, ACT 2901
Australia
auhicn4c@ibmmail.com

Michael J. Rothman
IBM
Hudson Valley Research Park
Hopewell Junction, NY 12533
rothman@fshvmdc.vnet.ibm.com

Abstract

This paper addresses the effectiveness of two data mining techniques in analyzing and retrieving unknown behavior patterns from gigabytes of data collected in the health insurance industry. Specifically, an episode (claims) database for pathology services and a general practitioners database were used. *Association rules* were applied to the episode database; *neural segmentation* was applied to the overlaying of both databases. The results obtained from this study demonstrate the potential value of data mining in health insurance information systems, by detecting patterns in the ordering of pathology services and by classifying the general practitioners into groups reflecting the nature and style of their practices. The approach used led to results which could not have been obtained using conventional techniques.

1 Introduction

The need to maximize the use of data for planning and strategic business development in all aspects of management has led many corporations to build comprehensive information systems that record all kind of operational transactions. As these databases grow larger, with gigabytes sizes becoming quite common, they are overwhelming the traditional query and report-based methods of data analysis [MPM96]. Data mining is the data-driven extraction of information from such large databases, a process of automated presentation of patterns, rules, and functions to a knowledgeable user for review and examination. A domain expert plays an essential role in the paradigm because he/she decides whether a pattern, rule, and function is interesting, relevant and useful [RM95].

Data mining techniques have been applied mostly to 'database marketing' through the analysis of customer databases; other applications include analysis and selection of stocks, fraud detection, spending patterns through the study of financial records, and detection of spatial patterns of bit failures in semiconductor memory fabrication [FPS96].

With the steady rise in health-care costs and the growing urgency to control these costs, timely analysis of health-care information has become an issue of great importance. Large corporations, hospitals, health-care maintenance organizations, and insurance companies require expert analysis of their health-care data, a task that is both time consuming and very expensive [MPM96]. Consequently, these organizations are developing automated mechanisms to support the process. For example, a system for detecting health-care provider fraud in electronically submitted claims has been developed at Travelers Insurance [FPS96].

The Australian Health Insurance Commission (HIC) has collected detailed claims information and has established a homogeneous claims database; this has been done through the administration of various programs.[1] The on-line claims file alone is over 550GB containing five years of history. Great effort has been directed towards the *accuracy* and *relevance* of the data, which are used in the development and implementation of strategies to detect and prevent fraud and inappropriate practice.

At HIC, analysis of claims data is performed by identifying and establishing claiming patterns

Proceedings of the 22nd VLDB Conference
Mumbai (Bombay), India, 1996

[1] HIC administers five major programs: Australia's Medicare, Pharmaceutical benefits scheme, Child care and rebate scheme, Medibank private, and Fraud and inappropriate practice prevention.

and relationships which represent appropriate utilization of services. To date, most data analysis has focused on the area of detection and prevention of fraud and inappropriate practice. Inappropriate practice deals with issues such as requesting or providing services which are unreasonable, unnecessary or excessive (e.g., indiscriminate ordering of cholesterol test). Typically, this type of analysis relies on human experts, but these experts are both expensive and scarce. HIC has supported the process with a neural network, though this approach can only be used on a subset of the database.

In addition, HIC has developed and implemented a pattern-recognition technique for classifying the degree of appropriate practice of service providers. By screening with this technology the claiming patterns of all practitioners, a score or class is assigned to each practitioner; those practitioners who have a high likelihood of an inappropriate practice pattern are identified. However, this technique allows only four different classes.

With the motivation of evaluating and determining the feasibility of using newer technology that could be applied to large databases, we addressed the effectiveness of two data mining techniques for analyzing and retrieving unknown behavior patterns from gigabytes of data collected in this health insurance information system. Specifically, an episode (claims) database for pathology services and a general practitioners database were used; *association rules* were applied to the episode database; *neural segmentation* was applied to the overlaying of both databases.

The results of this study have demonstrated the potential value of data mining in health insurance information systems, by detecting patterns in the ordering of pathology services, and by classifying the general practitioners into groups reflecting the nature and style of their practices. The approach used led to results which could not have been obtained using conventional techniques.

2 Data mining techniques applied

There are several primary data mining techniques including classification, regression, clustering, summarization and dependency modeling, among others [FPS96]. In this paper, we report on the applicability of *association rules*, which is a more sophisticated method of *summarization*, and *neural segmentation* as a particular implementation of *clustering*. These techniques, whose basic characteristics are summarized next, were selected based on the description of the problem and on our interest on experimenting with two distinct methodologies.

2.1 Association rules

Association rules over basket data type transactions is a problem defined in [AIS93]. The mining of association rules relates to finding intra-transaction patterns and can be defined as follows: given a database of transactions, where each transaction represents a set of items (e.g. services rendered), generate all associations such that the presence of some specific item(s) x in a transaction implies the presence of other item(s) y. The association rule $x \Rightarrow y$ will hold given a *support* and a *confidence* greater than a user-specified minimum support (s_{min}) and minimum confidence (c_{min}), wherein

support is the number (or fraction) of the transactions that contain a given item set.

confidence measures the frequency that items in a multi-item set are found together.

The database could be a data file, a relational table, or the result of a relational expression [AS94]. The strengths of this technique are the capability to handle large databases in an efficient manner, while its execution time scales almost linearly with the size of the data. These characteristics have proven true during the course of this study.

2.2 Neural segmentation

Neural segmentation is a pattern detection algorithm, in which the base technology is a self-organizing feature map [Koh89]. Self-organizing feature maps, otherwise known as topological feature maps, loosely preserve the topology of the multi-dimensional space in the two-dimensional map. That is, similar prototypes are near each other on the map.

A self-organizing feature map consists of a two-dimensional array of units; each unit is connected to n input nodes, and contains a n-dimensional vector w_{ij} wherein (i,j) identifies the unit at location (i,j) of the array. Each neuron computes the Euclidean distance between the input vector x and the stored weight vector w_{ij}. The neuron with the minimum distance is declared the "winner" and the input vector is assigned to this neuron. In addition, each of the weight vectors is modified as follows:

$$\textbf{\textit{new weight vector}} = w_{ij} + \mathrm{LR} * \mathrm{NF} * (x - w_{ij})$$

wherein

- ***LR*** is the *learning rate*, a linearly decreasing scalar which changes after each epoch;
- ***NF*** is the *neighborhood function*, a Gaussian distribution function in map-space, centered on the winning neuron;
- ***epoch*** is the presentation of all input vectors to

age	lab_id	benefit	cost_ser1	cost_ser2	....	service 1	service 2	....	provid.	sex

Figure 1: Record in episode database

provi	sex	tot_serv	%female pat	total patients	nursing vis	home visits	tot.benefit	cost/serv	..

Figure 2: Record in general practitioners database

the system once, that is, one iteration over the data.

Key to the success of this analysis is the representation of behavioral data. Since self-organizing feature maps are a form of clustering, care must be taken into properly balancing the inputs. That is, each vector element intrinsically represents one equally weighted dimension of the subject's behavior. Balancing the inputs implies that each equally-important aspect of the problem has the same number of vector elements. This is accomplished through an extensive study of the problem, with attention to the type of characteristics that are important to the solution, and to a careful review of the variance of each of the inputs.

The output of the algorithm is a two-dimensional array of *segments*, each one described by both its behavioral profile and potentially an overlay of further descriptive data.

3 Description of experiments and results

In this section, we describe the experiments conducted using association rules and neural segmentation. We also describe the database structure used and the corresponding results.

The experiments were performed using non-synthetic de-identified data obtained from the health insurance provider, which included the pathology episode and the general practitioners databases. These experiments were performed on an IBM System 9121 model 480 with 3380 DASD storage subsystem, using a single CPU (actual capacity of 19 MIPS). The data was stored in relational form within DB2/MVS.

3.1 Data description

The data used in these experiments is organized as two databases:

- A transactional database corresponding to historical events for a set of individuals, identified by physician number;
- An aggregate database from several other databases, including claims by service containing more than 18,000,000 records in one year. This database is used in other business decision applications.

Episode database

The episode database for pathology services is stored in relational form. This database contains 6,800,000 records and 120 attributes (3.5 GB). The records, which were sorted by provider identification, have the layout depicted in Figure 1.

General practitioners database

The general practitioners database contains 105 attributes and consist of approximately 17,000 records, which correspond to active general practitioners during one year period. By doing the aggregation, measures are obtained which assess the cost, usage, and quality of the services (e,g., big excisions, cancer removal, pathology benefit per patient); additional descriptive elements include data such as age or sex of the physician. This database was sorted by provider identification, and the record layout is shown in Figure 2.

3.2 Experiment 1: Association rules

The association rules algorithm was applied to the episode database for pathology services; each patient visit is mapped to a record in the database. Therefore, a database tuple corresponds to a unique identifier and one or more medical tests (or services) taken at a given instance in time, with a maximum of 20 tests per episode. There are three major steps involved when applying association rules:

- Data preprocess
- Application of association rules
- Postprocess or analysis of results

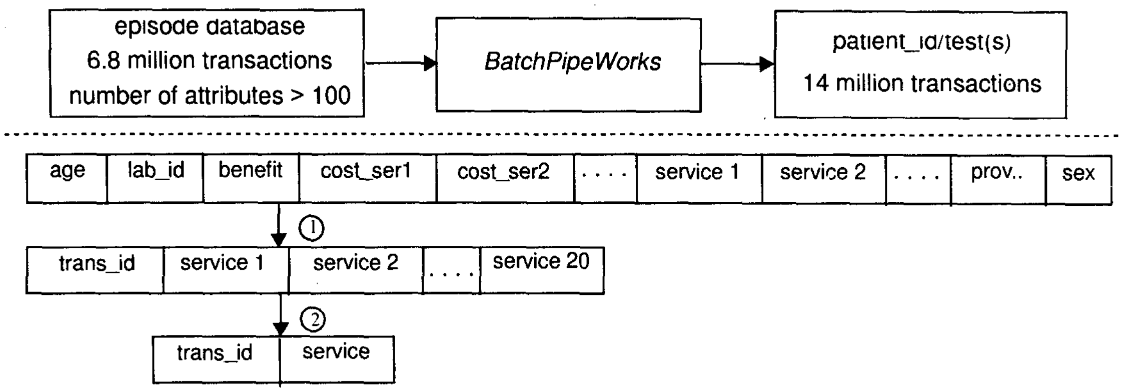


Figure 3: Data extraction for association rules

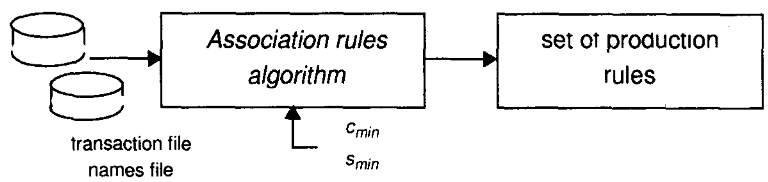


Figure 4: Application of association rules

Data preprocess

A preprocess phase was performed to extract from the database only those attributes to which associations would be applied. This process, depicted in Figure 3, relies on the utility BatchPipeWorks[BAT95] which allows manipulating millions of records in an efficient manner.

In this study, the *interesting* attributes are the pathology services (or medical services); the record format required by the association rules algorithm requires a transaction id and n-attributes (step 1). Since the schema definition for the episode database contains a maximum of 20 interesting attributes per event, pivoting was done such that the records were organized as a transaction-id/attribute pair (step 2). During the preprocessing, the unique transaction-id was based on the patient-id and time stamp, and it was added to each record while extracting the attributes from the database.

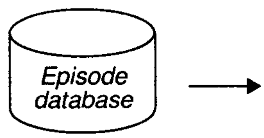
Application of association rules

Association rules were obtained by using the Apriori algorithm presented in [AS94]. The inputs include the transaction file described above, and a names file containing a text description for each code used in the transaction file, as shown in Figure 4. Minimum confidence c_{min} and minimum support s_{min} values were specified during the experimentation.

3.3 Experiment 2: Neural segmentation

Neural segmentation was applied to the episode database for pathology services. The aim was to examine the ordering profiles of general practitioners (GPs), and determine segments in the profile population by nature of the tests that the GPs ordered. The GP database was used for obtaining data describing the nature of the practice as well as the identification of the selection and frequency of tests. The records of 10,409 de-identified GPs were examined; these consisted of 6,817 vocationally registered GPs (65.5%) and 3,592 non-vocationally registered (34.5%). The data set contained 11.6 million tests ordered by these GPs. There were 227 different tests.

The episode database was used for creating the input vectors to the segmentation algorithm. Given that the records in the episode database contain up to 20 tests per episode, a rotation of the records and aggregation by GP is required. 10,409 vectors were created, which correspond to a scaled



n = 10,409
m = 227

	test 1	test 2	test 3	... test j	... test m
v_1	1	1	0	0	0
v_2	0	0.03	0	0	0
v_3	1	0.02	1	0	1
...	...	...	...	...	...
v_i	0.01	0	1	0.01	0
...	...	...	...	...	...
v_{n-1}	0	0	0	0.3	1
v_n	0	0	0	0	0.5

Figure 5: Neural network input vector

representation of the frequency of each test ordered by the GP during a one year period. The number of tests performed by each physician was scaled from 0 to 1 with respect to the total number of tests. The final format of the input vectors is depicted in Figure 5.

The system was trained iteratively by presenting the vectors v_i (with 1 <= i <= 10,409) for 25 epochs or iterations. The parameters used were

- learning rate (LR) with an initial value of 0.8 and a final value of 0.05; the learning rate was decreased linearly after each epoch;
- neighborhood function (NF) with the width of the Gaussian distribution function varied from the square root of the number of nodes to 0.1.

3.4 Results

Experiment 1

Association rules were obtained using c_{min} = 50% (minimum confidence), and three different values for s_{min} = 1%, 0.5%, and 0.25% (minimum support) specified by the user. Since there is a total of 6.8 million of episodes for pathology services (or transactions in the database), s_{min} = 1% represents 68,000 transactions. The number of association rules obtained in each of the experiments is shown in Table 1. Simple as well as complex rules were determined by the algorithm; an example of such rules is as follows:

*'If **Iron Studies** and **Thyroid Function Tests** occur together then there is an 87% chance of **Full Blood Examination** occurring as well. This rule was found in 0.55% of transactions.'*

Table 1: Test using episode database

	s_{min} = 1%	s_{min} = 0.5%	s_{min} =0.25%
c_{min} = 50%	24 rules	64 rules	135 rules

The first experiment was set to identify those pathology services which appeared in various combinations with s_{min} > 1% (or 68,000 transactions). A minimum confidence c_{min} = 50% was set to summarize the data into 24 production rules. The major results are:

1. The most frequently medical test (*test-A*) occurs in 28.15% of all episodes. An additional 20.15% occurred in association with a collection fee. This test was present in 48.30% of all episodes for pathology services.
2. The most commonly ordered combination of tests occurred together in 10.9% of episodes. The rules obtained show that there was a 62.2% chance that if *test-A* was ordered then *test-B* would also be ordered.

The second experiment was set to identify those services which appeared in various combinations with s_{min} > 0.5% (or 32,000 transactions). c_{min} = 50% was set to reduce the data into 64 production rules. A greater amount of knowledge through the behavior patterns was gained by setting s_{min} > 0.5% rather than s_{min} > 1%.

Lowering s_{min} to 0.25% (or 16,000 transactions), with retaining c_{min} = 50% produced 135 rules and information which was of much greater value, as demonstrated by the following two examples:

1. If *test-B* was claimed with *test-C*, there was a 92.8% chance that a *test-A* would also be claimed. This finding may be indicative of different ordering habits for similar clinical situations.
2. If *test-D* was ordered with *test-E*, it was revealed with a 60% chance that *test-A* would also be ordered, in 0.25% of the cases. This rule raises questions on whether this testing is of a screening nature, or even if this is the most proper medical monitoring regime.

The various examples above show commonly ordered tests in association in a relatively large number of episodes. The appropriateness of ordering combinations such as the ones described here needs to be considered by the relevant medical learned bodies. It would appear from the analysis carried out during these experiments that a large proportion of pathology tests ordered are of a non- specific screening nature rather than a planned approach.

Association rules have given a new perspective to this problem, and the outcome of more detailed data mining could provide the means for generalized and targeted education in conjunction with medical learned bodies. New and unexpected results obtained on *test-X* (as discussed below) demonstrates the value of association rules in uncovering unusual associations which are otherwise missed. A new higher-level perspective was gained about the ordering behavior of providers than what could be readily gained from conventional query techniques.

Effect of noise in the data

Various combinations of collection-fee items in combination with medical services create several non-meaningful rules and noise in the results; they obstruct significant findings and complicate the analysis of the output. Since 30.20% of all episodes contain a collection-fee item, an experiment was performed excluding from the pathology episode database all such items.

Filtering the collection-fee items produced much greater clarity in the association rules. At s_{min} = 1% there were only nine rules, in contrast to the 24 rules obtained previously. These new rules provided a high-level view of the ordering behavior in the services as well as in the billing practices.

An unusual combination relating to *test-X* became immediately apparent from the association rules after filtering the collection-fee items. The financial incentive to claim *test-X* rather than *test-Y* is $13.55 extra benefit per test. There is no restriction on claiming both items, presumably due to the very exceptional clinical possibilities of doing these two tests together. However, there can be little doubt from the results that *test-X* is being incorrectly claimed instead of *test-Y* in a large proportion of cases. There was an overpayment of more than $550,000 for these items in just one state during one year. The question arises on whether the same claiming practice exists in other states to the same extent. This problem has been present for over five years and had gone undetected by the monitoring techniques currently in place.

Execution time

Table 2 shows the relative execution times obtained when applying association rules to the episode database.

Table 2: Execution time

s_{min}	CPU time	Elapsed time
1.00 %	22.23 (min)	35.6 (min)
0.50 %	22.53 (min)	35.6 (min)
0.25 %	22.56 (min)	35.8 (min)

Experiment 2

The experiment used a map of 4x4 units with vectors of 227 elements. The size of the map was selected to provide a better classification than the one already available at HIC. The dimension of the vectors was determined by the number of tests. The neural segmentation process found prototypical patterns.

The number of segments determines the granularity of the description. In this method the number of segments is a user-defined input, and the method guarantees finding the segments specified. By applying our implementation of self-organizing maps to various business problems, we have found that there is no "correct" number of segments; moreover, the number of segments should be determined by the business purpose of the segmentation. In the experience described in this paper, the goal was to create an unbiased sub-division of GP practices for the purpose of more effective monitoring of test ordering. The 16 segments divided the GP population into groups with at least 100 in each, enough to get meaningful statistics on test ordering.[2]

For each segment, the results were tabulated according to the most frequently used test types in the database, ordered by relative interest (see Table 3).

After neural segmentation was completed using the episode database, we tagged each GP with the segment number of the closest node. This required computing the Euclidean distance between each GP's normalized test ordering record and each of the final weight vectors. We then merged the GP database with the marked GP ordering file, by GP ID. This added significant descriptive information to the analysis. Specifically, we computed the mean, standard

[2] A pragmatic matter must be noted: it is not possible to get meaningful segmentation in a true 227 dimensional space with a data set of 10,000 points. However, in our example the description is very sparse (many 0's). That is, a GP orders a set of tests that suit his practice, and there are really only few types of practice, therefore there is only a small number of dimensions.

Table 3: Test usage

Use	Average number of tests ordered by each member of the segment
Relative use (x)	"Use" divided by the average use for all GP's
Market	Fraction of a specific test ordered by a segment. Indicates the number of times the average test use a given segment displays. It shows the segment's relative interest
Penetration (pen)	Fraction of the population in this segment ordering at least one specific test

deviation, minimum, and maximum for each numeric quantity in the GP database, by segment ID. This included 105 different attributes such as the number of house calls, or nursing home visits, the age of the physician, and the fraction of female patients seen.

A summary description of the 16 nodes found is shown in Appendix A; the information derived for each segment by overlaying the attributes contained in the general practitioners database was added. Therefore, each segment represents the practice type assigned. The average neural network score[3] was also included in the segment description. Comments on each segment reflect the domain expert evaluation and interpretation of the data mining output produced.

From the grid depicted in Appendix A, it is clear that one large segment was created; this segment has a very low average of test performed per GP so it was qualified as a "low usage" segment. By overlaying the GP database, it was identified that the segment contained most of the non-vocationally registered GPs.

Neural segmentation enables the creation of as many nodes and subnodes as desired. Further segmentation was applied to the data constituting Segment 2 and Segment 4 to enable adequate analysis because those nodes were not densely clustered; four subnodes were produced for each one of these groups. In one of the nodes with an average neural network score of 1.8, the creation of four subnodes produced groups with scores between 1.4 and 2 which were a variation of the same practice style. In another larger and apparently less homogenous group, the creation of four subnodes produced some distinctly different groups of practice style. Breaking down this fairly heterogeneous sector further into four subnodes reveals quite distinct subgroups.

Applying neural segmentation to the general practitioners database successfully achieved the classification of general practitioners into groups of various sizes reflecting the nature and style of their practices. Meaningful categories have identified GPs with different medical characteristics (e.g., obstetricians, after hours clinics, environmental medicine, geriatric practices, etc.)

4 Concluding remarks

We have addressed the effectiveness of two data mining techniques when applied to large databases in a health insurance information system. Through this study, we have shown that data mining algorithms can be used successfully on *large, real customer data* (not synthetic data), with *reasonable execution time*. Moreover, we have also shown that these algorithms can result in *quantifiable benefits* for the interested organization, helping identify specific actions to be taken. In particular, among the results obtained we can mention the following:

- The study provided the health insurance organization with new and unexpected relationships among the combination of services provided by physicians in the area of pathology. An overpayment of over $550,000 in one year was uncovered; detection of this problem, which had existed for over five years, had eluded conventional monitoring techniques.
- The study provided a classification of general practitioners into groups of various sizes reflecting the nature and style of their practices. Categorization of general practice into its various subgroups had proven elusive in the past. The new subgroups will allow greatly improved monitoring of practice patterns. Given that the business goal of health insurance corporations is modifying the behavior of practitioners towards best practice, neural segmentation provides the means of understanding and monitoring the behavior of the various subgroups in an application such as general practice.

From this experience, we have empirically verified that large databases are indeed an important asset to organizations, which can use such databases to extract meaningful information with reasonable effort.

[3] The class or score was obtained from the classification of practice pattern using the neural network developed at HIC and trained by their domain experts.

References

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., May 1993.

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487-499, Santiago, Chile, September 1994.

[BAT95] IBM BatchPipes/MVS, BatchPipesWorks Reference Manual, September 1995.

[FPS96] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery: an overview. In *Advances in Knowledge Discovery and Data Mining, Chapter 1*. AAAI Press, 1996.

[Koh89] T. Kohonen. "Self-Organization and Associative Memory," Springer-Verlag, New York, 1989.

[MPM96] C. Matheus, G. Piatetsky-Shapiro, D. McNeill. Selecting and reporting what is interesting. In *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996.

[RM95] M. Rothman, E. Murphy: Data mining: A practical approach for database marketing. In *IBM White Paper*, Dallas, Texas, April 1995.

Appendix A - Segments

segment 0 97 GPs class 1.8 3.6% of total tests ordered 4347 average tests/GP Manage cancer patients with or without proper training & supervision	**segment 1** 206 GPs class 2.7 4.8% of total tests ordered 2733 average tests/GP GPs practicing in the country, doing less pathology and when done, these are often obstetrics tests	**segment 2** 102 GPs class 1.8 4% 4537 average tests/GP Extended hours clinic	**segment 3** 54 GPs class 1.3 2.7% 5736 average tests/GP Self professed experts outside mainstream (e.g., chronic fatigue)
segment 4 445 GPs class 2.4 8.4% 2204 average tests/GP Four sub-nodes: - unusual tests-hosp yg - unusual test-hosp old - unusual tests-"usual" GP practice - unusual tests-medical centers	**segment 5** 149 GPs class 1.7 4.8% 3745 average tests/GP "Classic" over servicing-bad	**segment 6** 151 GPs class 2.0 4.4% 3422 average tests/GP "Classic" over servicing-less bad	**segment 7** 442 GPs class 2.3 8.7% 2288 average tests/GP "Average GP" using acupuncture
segment 8 1735 GPs class 2.7 18.5% 1240 average tests/GP Older GPs with small practices	**segment 9** 127 GPs class 1.8 4% 3677 average tests/GP Environmental medicine	**segment 10** 227 GPs class 2.5 5.1% 2631 average tests/GP Young GPs in acute medical clinics	**segment 11** 26 GPs class 1.9 0.8% 3471 average tests/GP AIDS, STD, drug addiction clinics
segment 12 5450 GPs class 2.8 6.2% 133 average tests/GP Body of GPs represent conventional general practice (many part time)	**segment 13** 500 GPs class 2.2 11.2% 2604 average tests/GP Geriatric practice	**segment 14** 595 GPs class 2.9 9.2% 1802 average tests/GP Female GPs working on female health	**segment 15** 103 GPs class 2.2 3.7% 4166 average tests/GP Young female GPs in women's health clinics