# Feature Selection for Examining Behaviour by Pathology Laboratories

Simon Hawkins, PhD
Strategic Business Development Manager

Graham Williams, PhD
Principal Research Scientist
Enterprise Data Mining

Rohan Baxter, PhD
Group Leader
Enterprise Data Mining

CSIRO Mathematical and Information Sciences,
GPO Box 664, Canberra ACT 2601, Australia.

### Abstract

Australia has a universal health insurance scheme called Medicare which is managed by Australia's Health Insurance Commission (HIC). Medicare payments for pathology services, for example, generate voluminous transaction data relating patients, doctors and pathology laboratories. Systems are required to monitor the utilisation of pathology services which are currently growing faster than any other area of medicine except radiology. Systems currently exist to automatically monitor pathology-ordering behaviour by individual doctors. The next step is to derive a new set of features capable of characterising the time-varying nature of pathology processing by pathology laboratories. These can then be used as inputs to a predictive modelling system. The current study identifies a new set of features for characterising the time-varying behaviour of pathology laboratories in one state of Australia. These features are presented visually and then used to cluster similar pathology laboratories and to detect outlying laboratories with unusual behaviour. Data organisation and data transformation methods are described for the efficient access and manipulation of these new features.

**Keywords:** Health information management; data mining; feature selection; Pathology; decision support system; massive administrative database; inappropriate medical practice; Medicare.

# 1 Introduction

In the context of the knowledge discovery in databases (KDD) process,[1] the earliest steps of data pre-processing and feature generation are known as feature generation. The later steps of the KDD process involve pattern searching using data mining techniques, model evaluation and model deployment. The literature has primarily focused on these latter stages, because in many domains, features are considered both obvious and known a priori.[2] However, deriving a good set of features from administrative health claims databases is far from straightforward. Moreover, we have found good feature sets produce much greater performance gains in this domain than are achieved by trying different predictive modelling techniques. In the current study we report on our search to find a good feature set for characterising the pathology processing behaviour of pathology companies in the state of New South Wales (NSW).

In Australia pathology is a 1.4 billion dollar industry that is growing more rapidly (60 million dollars per year) than any other area of medicine except radiology. Also, doctors order pathology tests for their patients from private pathology companies. One or more tests are ordered together in groups called "episodes." These tests are almost entirely paid for by government funding under the universal health insurance scheme known as Medicare. Medicare payments for pathology services generate voluminous transaction data on patients, doctors and pathology laboratories. These payments are administered by a government agency called the Health Insurance Commission (HIC). The HIC's charter is to make accurate and timely payments while at the same time ensuring pathology test are appropriately ordered. The HIC wants to use this massive, online transaction database to monitor the behaviour of pathology laboratories.

The HIC has already developed capabilities for training and deploying artificial neural network models that detect individual doctors who are at "high risk" of practising inappropriately. The reason for this inappropriate practise is sometimes found to be excessive or indiscriminate pathology ordering. These predictive models currently use a small number of features as inputs. These features are based upon simple counts available from the online transaction processing system. These features typically include volumes of transactions, the types of pathology tests performed and the dollar value of tests performed each quarter for the pathology laboratories.

There is no current system, however, for examining the behaviour of pathology laboratories. Such a system may be worthwhile because groups of 'high risk' doctors appear to order their pathology from a small number of pathology companies. The major challenge in developing such a model is to develop a new set of features that are sensitive to unusual behaviour by laboratories rather than individual doctors. These features will be derived from the Medicare transaction database and will describe the processing patterns of pathology laboratories over time. These features will then be incorporated into predictive modelling techniques, such as neural networks and decision trees, in order to examine the behaviour of a pathology laboratory.

The remainder of the paper is organised as follows: Section 2 describes the data organisation and data transformations undertaken before features could be generated and visualised efficiently and flexibly. Section 3 describes the feature sets generated, some data mining methods that utilise these features and some visualisations of the feature sets. Section 4 describes additional features with time components that were investigated. Section 5 describes some prospective benefits in using the generated features. Section 6 presents our conclusions.

# 2 Data Organisation

In this section we describe the available data, the data transformations and the data organisation used to enable the required fast access for the feature generation methods.

## 2.1 Data Types

The project data were all pathology services paid for under Category 6 of the Medicare Benefits Schedule for the state of New South Wales over the eight quarters in 1997/1998. Table 1 summarises the dataset. Additional data on referring doctor attributes were also provided.

Each transaction has 44 fields relating to four distinct entities. They are the *pathology laboratory*, which performs the pathology test, the *doctor*, who orders the test, the *patient*, for whom the test is ordered and the *transaction* itself. Table 2 summarises the transaction fields. The 36.8 million transactions covered 79 pathology laboratories, $20,314$ doctors and $3,853,603$ patients. The HIC is required by law to de-identify fields that could identify any individual entity. This was done by encrypting entity identifiers and postcodes. In lieu of postcode location information, a coding scheme called RRMA (Rural, Remote and Metropolitan Areas) is used to group postcodes into seven different types of geographic regions, including rural, metropolitan and city.

## 2.2 Data Transformation

The following pre-processing was performed:

- The five date fields were converted to day offsets, starting from January 1, 1970 (the Unix epoch starting date). The offsets for dates before January 1, 1970 are negative. This simplified the calculation of time lags used in feature generation.

- Empty field values were replaced with a marker value.

- Since a different subset of pathology tests have item numbers changed each year, all test item numbers were mapped to those current at June 1999.

Pre-processing was performed using the Perl scripting language. Use of this language improved performance by an order of magnitude over other scripting languages such as Tcl. This performance improvement made the current study feasible. It requires just three hours for a Perl script to make a single pass over the entire data using a ten 167MHz-processor, 4.5 Gigabytes Sun 4000 Enterprise server. A Tcl script with the same functionality required 72 hours. Since many passes over the data are required during the data transformation and data stratification stages of the study a three day wait for each pass would soon take up a significant proportion of the project time.

## 2.3 Data Organisation and Access

The transactions were originally stored in a single large relational database table. One approach, and a current area of research in data mining, is to interface data mining methods with a relational database[3]. However, this is a sub-optimal approach in the present study. Our explorations of feature generation required fast access to one or more individual transaction columns, whereas a relational database only provides fast access to individual transaction rows.

Alternative approaches for fast column access include data-cubes[4] and sufficient statistic caching.[5] These approaches are efficient for specific data methods such as associative rules or clustering, but are not efficient enough for intensive exploration of interesting features. For these reasons, we developed a column-binary-flat file approach for feature generation that allows our feature generation programs to selectively access one or more columns in an efficient and flexible manner.

### 2.4 Data Stratification

We stratified the data into the four categories shown in table 3. The motivation for the stratification was two-fold:

- The subsets are smaller and more manageable for data manipulation.
- It was expected *a priori* that pathology ordering patterns within each subset would be distinct. This expectation was only partly borne out. We later found that test ordering patterns for the *go* and *so* subsets to be indistinguishable.

## 3 Features for Pathology Laboratory utilisation patterns

The purpose of this project is to find a new set of features for describing the patterns of pathology tests processed by individual pathology laboratories over time. These features will provide the input to a model that will monitor behaviour by pathology laboratories.

### 3.1 Relative volume of tests in each subset

The first feature generated is the relative proportion of tests processed by a pathology laboratory in each of the four test categories (”four category feature”). The relative volume of tests processed, rather than an absolute count of tests, permits comparison between laboratories of different overall size.

A k-means clustering method is used to group the 79 pathology laboratories into four clusters on the basis of this ”four-categories feature”. Laboratories in the first two clusters (containing 17 and 15 laboratories respectively) mainly process pathology tests ordered by specialists consulting patients both in- and out-of-hospitals (sh and so). Laboratories in Cluster One process about 65% of their tests for doctors in the sh category. Laboratories in Cluster Two process more than 60% of their tests for doctors in the so category. The 25 laboratories in Cluster Three (the largest cluster) almost exclusively process tests ordered by doctors for patients out-of-hospital (go and so). More than 60% of the tests processed by these laboratories are ordered by GPs; 30% are ordered by specialists. Cluster four contains the 22 laboratories that process more than 85% of their tests for GPs ordering for out-of-hospital patients (go).

## 3.2 Univariate Features

A simple but effective data mining method is to examine outliers. For univariate continuous features, an outlier is defined as a value that lies more than two standard deviations from the mean for that feature within each of the four categories of laboratories. For univariate categorical features, an outlier is defined as a relative proportion which differs from the mean by more than 10%.

Using the 44 transaction fields as features, we searched for outliers. A number of features were found to be discriminatory in the sense that they identified just a small proportion of pathology laboratories as outliers. These features related to the properties of doctors ordering from the laboratory, the range of patients being tested by the laboratory as well as the properties of the laboratory itself:

Discriminatory laboratory features:

- The relative proportion of tests over the 11 pathology test categories (which include the Chemical, Haematological, Microbiological, Cytological, Immunological and Cytogenics categories).
- Sudden changes in the total volume of tests processed by the laboratory over time.

Discriminatory doctor features:

- Doctor specialty.
- The relative proportion of doctors within each geographic region who are ordering tests from the laboratory. Location is given by a RRMA code. RRMA is a coding system that identifies seven geographic region ranging from Capital City to Remote Rural.

- Proportion of high-ordering doctors ordering from a laboratory.

- Proportion of doctors rated as "high-risk" who are ordering predominantly from this laboratory. These ratings are generated by the HIC's artificial neural network systems that risk-rate individual doctors according to the likelihood they are practising inappropriately.

- Doctors who have suddenly cease ordering from other laboratories and switched to this one.

Discriminatory patient features

- Patient gender.

- Patient's home country, as indicated on their passport.

- Outliers from mean distribution of patient's age distribution.

The interestingness of these results cannot be decided from the data alone and follow-up is required with the client. Some explanations will be structural. For example, it is well understood that those pathology laboratories with an unusual distribution of doctor's geographic region will have a particular geographic market niche. Moreover, doctors operating within particular geographic regions (such as inner cities, for example) will see patients with particular medical conditions (such as STDs and drug-related problems) and these will necessitate highly specialised pathology investigations.

# 4 Features with a time component

We describe some relevant temporal features and visualisations for characterising pathology laboratory behaviour.

We also generate an entirely new set of patient features that we call the patient's 'loyalty score', 'attrition score' and 'departure rate'. These features summarise the relationship over time between doctors and pathology laboratories and also between patients and doctors. These features will be described further elsewhere.

## 4.1 Doctors changing pathology laboratories

An important area of current data mining research is the development of techniques for visualising the results of time series analysis[6] and event sequence analysis[7]. We develop a technique for visualising the time-varying relationship between the ordering patterns of doctors and pathology laboratories. Since there are over 20,000 doctors in the data, it will be important to develop an automated method for identifying interesting relationships between doctors and pathology laboratories. This is done manually in the present study, with several examples presented below:

Figure 2 graphically represents the relationship that two doctors have with pathology laboratories over time.

Graphs on the left side display the time-varying pattern of test ordering by the doctor from all 79 laboratories in NSW. Each × symbol stands for any number of tests that the doctor has ordered from a specific laboratory on a given day. Graphs on the right side show the total number of tests per week (TpW) that this doctor has ordered from any laboratory in NSW. The combination of these two plots shows when, where and how many pathology tests a doctor has ordered. Notice the difference between the two doctors: The doctor in the upper two graphs orders pathology at a consistent rate from just two pathology laboratories. The doctor in the lower two graphs suddenly increases their test ordering at about the same time as they switch ordering to three new companies.

## 4.2 Service Lags

A *service lag* is the time interval between the date on which the pathology test is ordered by the doctor, (*date of referral or DOR*) and the date on which the laboratory processes the test (*date of service or DOS*).

The summary of service lags in Figure 3 reveals that most tests are processed within 6 months (180 days) of the referral date. This compares with 66,705 tests with a service lag of six to 12 months and 13,827 tests with a service lag of more than one year.

Outlier detection using the service lag feature reveals that three laboratories have much longer service lags than other laboratories. Whilst unusual service lag distributions for a pathology laboratory may be due to the nature of the pathology tests performed by that laboratory, it may also indicate unusual claims behaviour that is worthy of closer examination.

### 4.3 Patient Episodes

A doctor orders a group of pathology tests to elucidate the nature of a medical condition. These tests make sense as a group (called an episode) but not on an individual basis. An *episode* is defined as the set of pathology tests ordered by a particular doctor for a particular patient on the same day.

Various test-based features of episodes are examined. Laboratories are identified that have more repeated tests in an episode than is typical. The duration of episodes is also considered. Approximately 6.5 million episodes were initiated in 1997. For 99% of these episodes, all tests in the episode were performed by a single laboratory.

Using episodes as features complicates the analysis because it introduces windowing effects. Episodes that are initiated before the beginning of 1997 continue into 1997. Episodes that start near the end of 1998 do not end until after the available data window. Care must be taken to ensure that these incomplete episodes do not skew results.

Figure 4 presents the distribution of the number of tests in each episode. The majority of episodes contain between 2 and 4 tests. This not surprising given that the HIC will only reimburse the three most expensive tests in an episode for non-hospitalised patients. In larger episodes, we note that episodes of size $2k$ are much less frequent than episodes of size $2k+1$. There seems to be a sharp break point in the number of episodes containing 60 or more tests.

In figure 5 we see the distribution of episode duration. This is the time taken to process all tests in an episode. The duration of the episodes seem to fall into one of several categories: zero days, up to a week, less than half a year and less than a year. As expected, there are no episodes with duration longer than 2 years, given the 2 year span of the data set.

## 5 Measuring Benefits

As one would expect, some of the pathology laboratory utilisation patterns discovered from the new features presented here are known from alternative sources of knowledge, while others are novel. The features generated in this paper should be used in new predictive models and their performance compared to existing methods. Only then can we determine whether they have led to any changes in policy regarding the payment of claims. However, the HIC has already commenced development of a web-based software tool that allows medically trained HIC staff to examine the behaviour of pathology laboratories on the basis of the new set of features identified in this study.

Of longer term interest is whether pathology ordering patterns can be compared against a standard of best practise. The is problematic because the Medicare claims dataset provides no information (such as clinical diagnosis) as why the doctor ordered the pathology tests. In some cases, clinical diagnoses might be inferred from other administrative data. For example, there is a standard battery of tests for the second trimester of pregnancy, and so differences from the standard, through under- or over-servicing may be observed. It is an open question whether this type of inference can be usefully obtained from administrative claims data. For instance, over-servicing can be confounded with further co-morbidity investigations.

# 6 Conclusion

Our investigation of features with a time component, such as service lags and patient episodes, utilised summary statistics and visualisations. The features are judged to be interesting using a detailed manual interpretation of similarities and interesting outliers. Algorithms for automating the process of finding outliers and for clustering entities characterised by features involving multivariate time series and outliers are needed in this domain and are not currently available.

It was a weakness of this study that closer iterative interaction with the client was not always available to direct the search for interesting utilisation patterns. However, this problem is common in data mining. Knowledgeable domain experts are in demand and are difficult to access. We recommend that attention and resources be paid to this issue in the structure of a data mining project.

Feature generation is an important step of the KDD process that has not appeared prominently in the literature. In the current project, we identify a number of new interesting features that may prove useful in a future predictive modelling system. These features were summarised, visualised and used as inputs for clustering and outlier detection methods. Data organisation and data transformation methods were described for the efficient access and manipulation of these new features. Further work is required for feature selection and training of predictive models with the new features and evaluation of performance against the currently deployed models.

# References

1. Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery:An Overview. In Advances in Knowledge Discovery and Data Mining, Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, 1–36. AAAI Press, Menlo Park,CA. (1996).

2. Liu, H. and Motoda, H. *Feature selection for knowledge discovery and data mining.* Kluwer Academic Publishers, Boston, (1988).

3. John, G. H. and Lent, B. Sipping from the data firehose. In Third International Conference on Knowledge Discovery and Data Mining, Heckerman, D., Mannila, H., Pregibon, D., and Uthurusamy, R., editors, 199–202. AAAI Press, Menlo Park,CA., (1997).

4. Agarwal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J. F., Ramakrishnan, R., and Sarawagi, S. On the computation of multidimensional aggregates. In Proceedings of the 22nd International Conference on Very Large Databases, Vijayaraman, T. M., Buchmann, A. P., Mohan, C., and Sarda, N. L., editors, 506–521 (VLDB, Mumbai (Bombay), India, 1996).

5. Moore, A. et al. Cached sufficient statistics for automated mining and discovery from massive data sources. Technical report, Robotics Institute and School of Computer Science, Carnegie Mellon University, (1999).

6. Keogh, E. and Smyth, P. A probabilistic approach to fast pattern matching in time series databases. In Third International Conference on Knowledge Discovery and Data Mining, Heckerman, D., Mannila, H., Pregibon, D., and Uthurusamy, R., editors, 24–30. AAAI Press, Menlo Park,CA., (1997).

7. Agrawal, R. and Srikant, R. Mining sequential patterns. In Proceedings of the 11th International Conference on Data Engineering, Yu, P. S. and Chen, A. L. P., editors, 487–499 (ICDE-95, Taipei, Taiwan, 1995).

| Period | File Size | Transactions |
|---|---:|---:|
| Quarter 1, 1997 | 680MB | 4,448,547 |
| Quarter 2, 1997 | 704MB | 4,529,848 |
| Quarter 3, 1997 | 706MB | 4,496,426 |
| Quarter 4, 1997 | 700MB | 4,423,777 |
| Quarter 1, 1998 | 749MB | 4,777,493 |
| Quarter 2, 1998 | 730MB | 4,640,190 |
| Quarter 3, 1998 | 758MB | 4,819,261 |
| Quarter 4, 1998 | 733MB | 4,623,801 |
| Total | 5.8GB | 36,759,343 |

Table 1: Number of transactions by quarter.

| Entity | Entity fields |
|---|---|
| Transaction | Test item number, date of service, date of processing, date of referral, date of lodgement, schedule fee for test, benefit paid, hospital indicator |
| Pathology Laboratory | unique identifier, postcode, RRMA |
| Doctor | unique identifier, postcode, RRMA, Specialty (GP or specialist) |
| Patient | unique identifier, date of birth, gender, postcode, RRMA, home country, age |

Table 2: Summary of Transaction fields, grouped by the Entity they describe.

| Test subset | Code | Description |
|---|---|---|
| Specialist, in hospital | *sh* | Test ordered by specialist doctor for patient in hospital. |
| Specialist, out of hospital | *so* | Test ordered by specialist doctor for patient out of hospital. |
| GP, in hospital | *gh* | Test ordered by General Practise(GP) doctor for patient in hospital. |
| GP, out of hospital | *go* | Test ordered by GP doctor for patient in hospital. |

Table 3: The four subsets of the stratified data

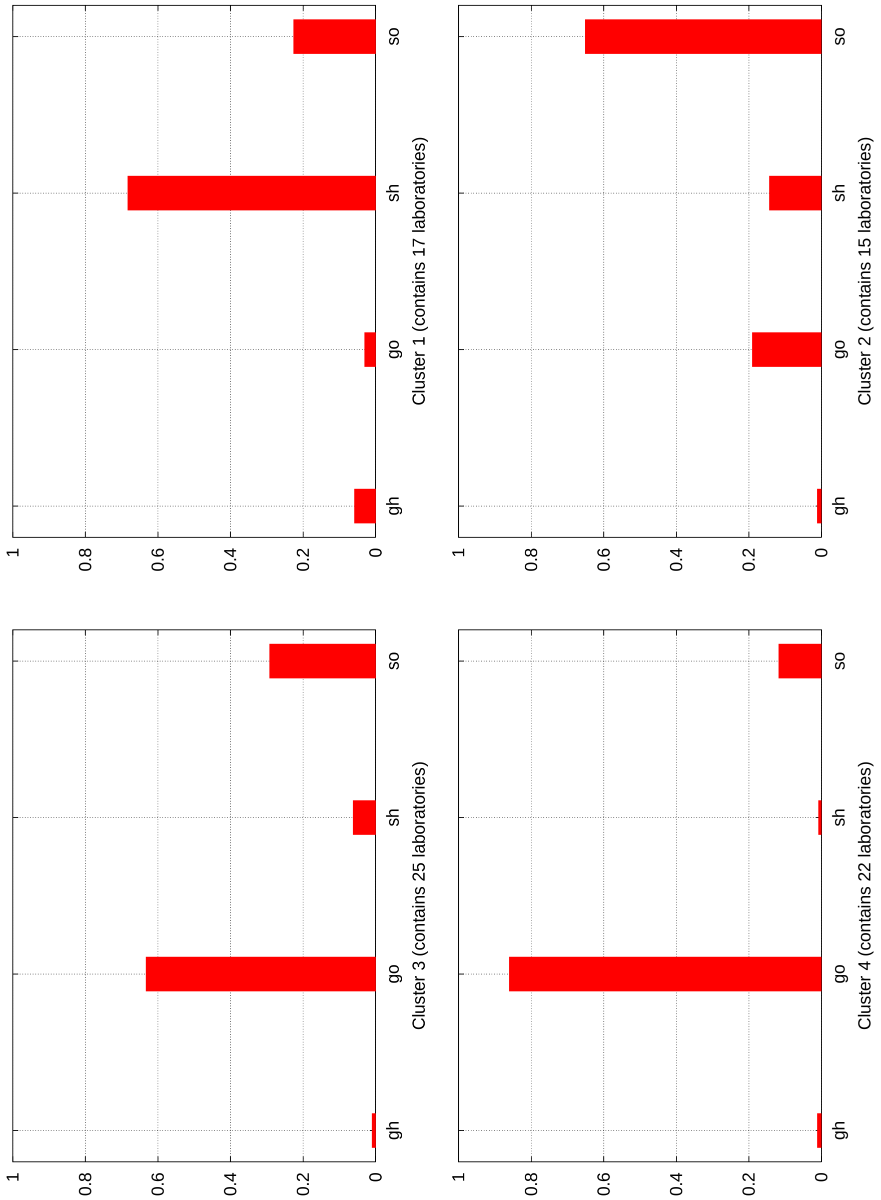

Figure 1: Laboratories clustered according to relative volume of tests for each doctor subset. For each cluster, the relative volume of tests in the four groups `gh`, `go`, `sh` and `so` is given.

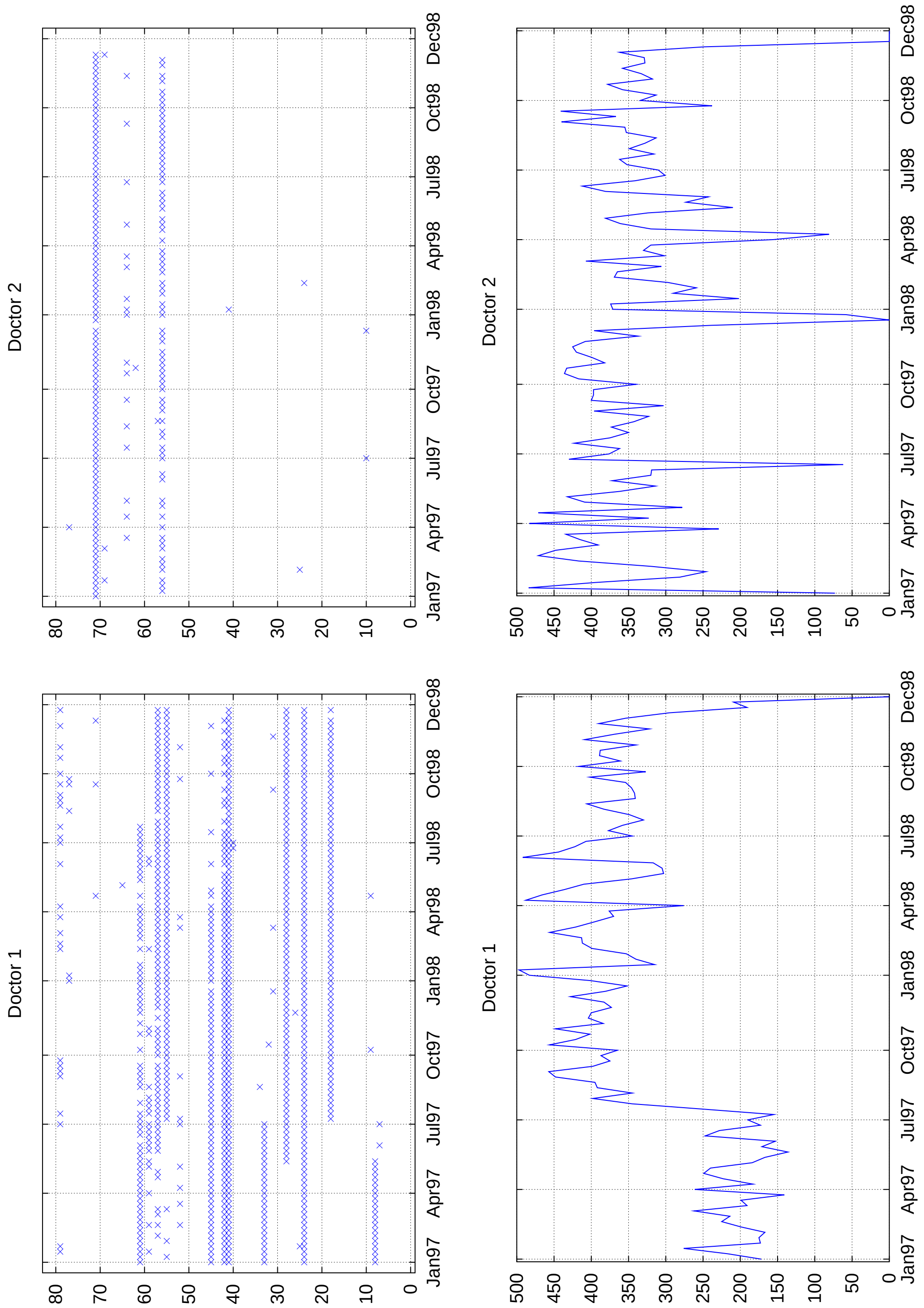

Figure 2: Pattern of laboratory use and tests per week for two doctors

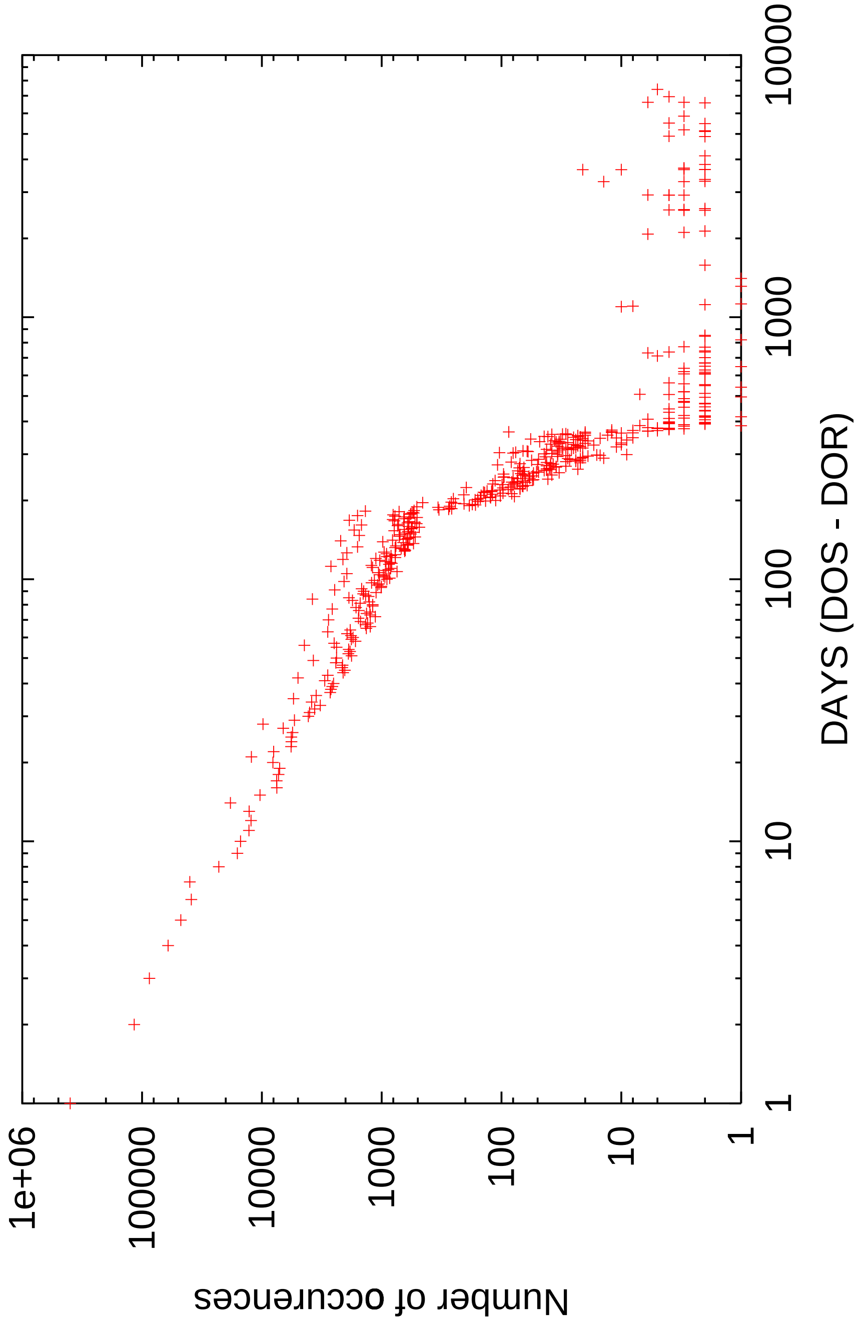

Figure 3: Service lags (time interval between `DOR` and `DOS` against number of tests performed during each interval). Two break points between main time intervals fall approximately at 183 days (6 months) and 365 days (1 year).

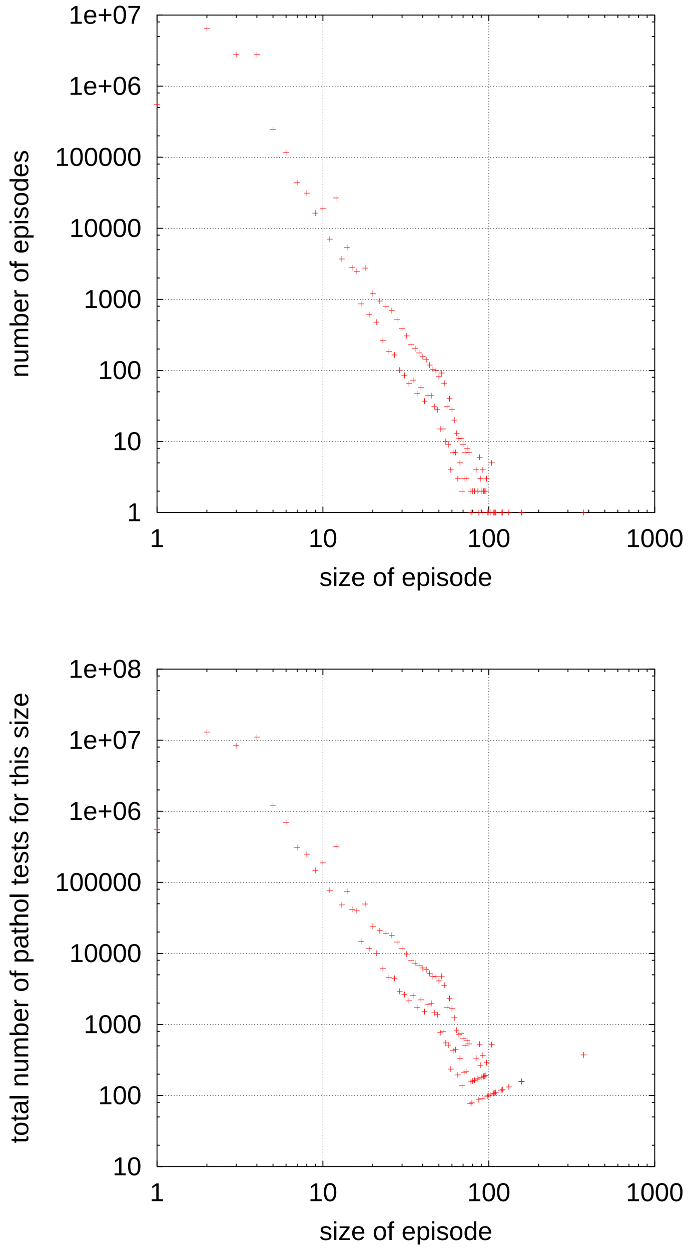

Figure 4: a) Number of episodes depending on their size b) Number of pathology tests for all episodes of a certain size

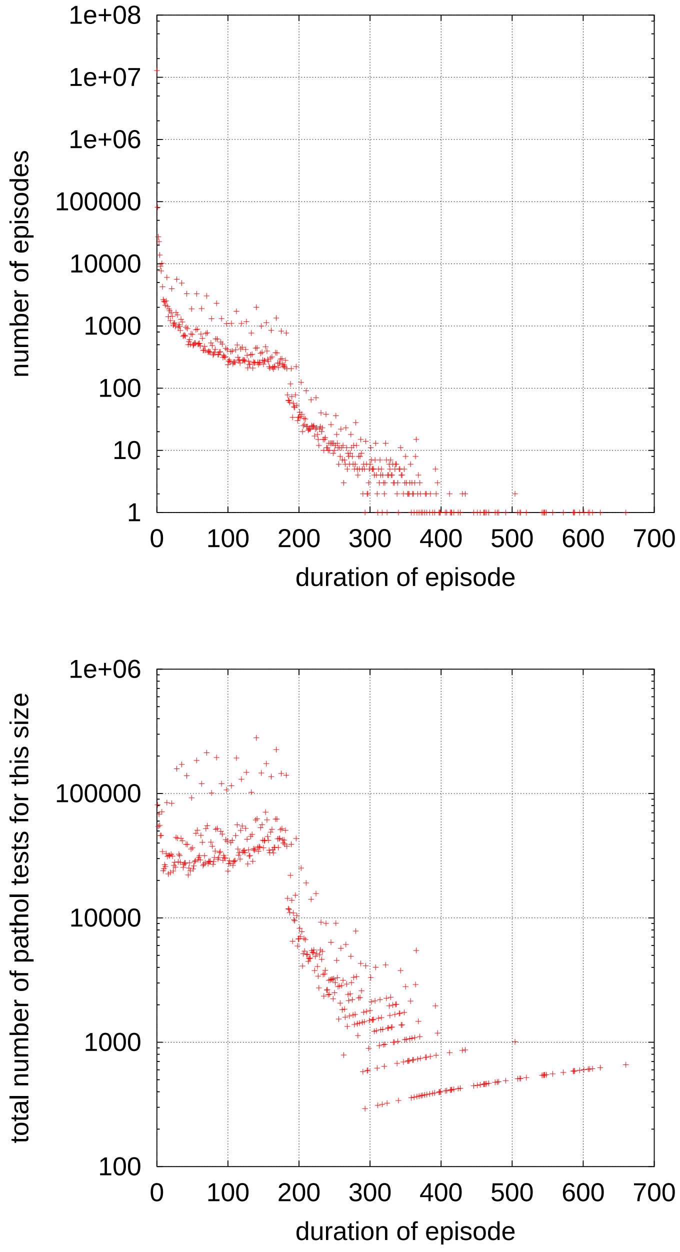

Figure 5: a) Number of episodes as a function of their duration b) Number of pathology tests for all episodes of a certain duration