electronic Journal of Health Informatics
http://www.ejhi.net
2006; Vol 1 (1): e2

Finding Anomalies in Medicare

Robert A. Pearson, Wayne Murray, Thomas Mettenmeyer
Program Review Division, Medicare Australia,
Tuggeranong, ACT, Australia

Abstract
The general behaviour of a medical practitioner can be assessed by metrics related to the services rendered. The values of these variables are inter-related and also related to the demographics of the practice. If there is a functional form that relates specified metrics to the others and the demographics, machine learning techniques can be used to determine the function. Both boosted regression trees and feedforward neural networks trained with backpropogation are used to learn a functional form. Where the predicted value of the function does not match the actual one, the behaviour is anomalous. This may be because there is fraud. Another technique for determining atypical behaviour is to reconstruct a value from the significant principal component scores, and compare this to the actual ones. Self organising maps can also be used to find similar, and dissimilar patterns. Both the targets themselves and the Karhunen-Loeve transform are used to train self organising maps. The coordinates on these maps are used to compare the doctors identified by the various techniques.

Keywords: Machine Learning, Atypical Behaviour, Neural Networks, Boosted Regression Trees.

1 Introduction

The Health Care system is a significant part of the national economy. In Australia, in 1995 it was about eight percent of the gross domestic product; some of the costs of the system are paid by the patients themselves; some are paid by Health Insurance funds, who in turn obtain the money from the people who insure themselves; some are paid by government bodies, and these are obtained by taxation. Most of the fees relate to justifiable services but some fraud occurs. In Australia there are various estimates of the actual amount of fraud in the Medicare system. The potential cases of fraud are identified through a number of techniques. One method is to examine the data relating to the health service and identify cases for further examination by subject matter experts. Various data mining techniques are applied. One way to classify these techniques is that some are based on traditional statistical techniques, others use machine learning (Artificial Intelligence) approaches.

Within both these groups some analyses use cases of known fraud to identify other cases as suspicious. The data in the suspicious cases have similar values of the independent variables and thus the dependent variable (fraud) is likely to be the same. This group of techniques has at least two disadvantages. Novel, different, or new ways for defrauding the system cannot be found, and the cases which are known to be of interest are only a small percent of the whole. This latter disadvantage can be ameliorated by using data transformation and utilising techniques such as support vector machines.

The techniques we use highlight providers whose behaviour is different from their peers, rather than fraud, obtaining benefits deliberately by deception, or abuse, which is unnecessary or inappropriate servicing. It is the responsibility of the subject matter experts to determine if the unusal consitutes genuine fraud or abuse.

Another approach is to assume that the fraud is reflected in the data as a case which is different from the other, more frequent practices. However a case that is different may be due to legitimate differences. This approach can be considered as two different groups of techniques. One uses some form of supervised learning. The other uses some way of clustering the data. Supervised learning considers that a number of variables, the 'targets', are functions of other variables. These may include as independent variables some of the other 'targets'.

All techniques produce cases where fraud has occurred but is not identified (false negatives) and cases where potential fraud is predicted but is not present (false positives).

Within Medicare Australia two different types of data are used for examination. One uses episodes of care, the other considers data related to a particular provider over a quarter. Use of an episode of care approach can be found in the literature [Copland, 2004, Semenova et al 2004]. This paper uses data related to the medical services of

The *electronic Journal of Health Informatics* is an international journal committed to scholarly excellence and dedicated to the advancement of Health Informatics and information technology in healthcare. ISSN: 1446-4381

a provider over a quarter.

Episode of care is the time period used in other technical analysis used within Medicare Australia. The decision to base the analysis on quarterly data rather than yearly data, was also made already. Whilst it is true that seasonality may have an influence, we have discovered that the demographics of the patients also change seasonally. The subject matter experts review the practitioners rolling year's worth of data before reaching a conclusion

2 Methods

The data

The original data source is the record of every service of a provider. This includes the identity of the patient, the amount charged, the refund due, facts about referrals, and diagnostic services, and pathology. These are aggregated over a quarter. From the patient details the proportion of the services to a gender age group range are calculated. There are twelve groups, one for each gender in an age group range of fifteen years. Some variables measure the totals, these include the number of services and amount of benefits paid. Other variables are fractions. These include means over items or unique patients. Six identify the fraction of patients who receive each of a number of services. Ten fractions relate to the number of requests for frequent items, five for the items requested, five for the rendered ones. Of these metrics relevant to a provider over a quarter, twelve are chosen as 'targets'. Some of the target variables, such as the benefit, are totals for a quarter, others are ratios (e.g. 'avrenbpat', the mean benefit per unique patient). Some of the original values used in the calculations appear in multiple targets. Some of the 'targets' are ratios that include other targets, sometimes as the numerator but more commonly as the denominator.

The information on the provider includes an identification number, the speciality and gender. The state and postcode of the provider's principle practice are recorded. The identification id and the postcode are not used in any of the analysis.

The results described in this paper are for the second and 3rd quarters on 2004. Only doctors in general practice were included. The Neural Network approach is least accurate when there is a long tail on the distribution of the data, and the Boosted Regression trees can be sensitive to outliers, it uses the mean which is not robust. For these reasons the tails of the distributions were kept for separate analysis. For the 2nd quarter the extremes were when the difference from the mean was three estimates of the standard deviation. For the 3rd quarter those excluded were more than four times the standard deviation from the mean at the high end and three at the low end.

While twelve demographic groups are recorded there are only eleven independent ones. The demographics are proportions of the total number of patients and as such sum to one. A principal component analysis was conducted and the scores that correspond to the first few important eigenvectors, and their eigenvalues were used in the subsequent analysis. Principal Component Analysis was used for data compression, to reduce the hyper-dimensionality of the data without sacrificing its explanatory power. This is included in the paper. Initially four scores were retained, now six are to be used. For the 2nd quarter of 2004 four cover 72% of the variance, while for the 3rd quarter they cover 68%. Six cover 82% and 80%. The proportion of the variance explained by the sixth eigen value is .043 for the 2nd quarter and 0.055 for the 3rd quarter. The analysis was performed using the NAG library©. The matrix used for the singular value decomposition was the correlation matrix. At first the scores were standardised for the unit matrix, but the current normalisation is that the variances are equal to the corresponding eigenvalue. This allows simple calculation of scores corresponding to the eigenvectors of a quarter. Unless specified otherwise all results relate to the 3rd quarter.

Table 1: Target Variables used in this analysis

Variable Name	Explanation
AVREFBITM	Average referred BENEFITS per ITEM
AVRENBITM	Average rendered BENEFITS per ITEM
BENEFITS	BENEFITS paid to the provider by Medicare Australia
DISER	Number of Diagnostic imaging services
ITEMS	Number of unique ITEMS rendered
PATHSER	Number of Pathology SERVICES
REFPATS	Number of Patients referred
REFPROVS	Number of unique providers to whom patients were referred
REFSERVS	Number of SERVICES referred
SERPATREF	Number of SERVICES performed on referred patients
SERPATREN	Number of SERVICES provider has performed on patients
SERVICES	Number of SERVICES

Target variables

To an accuracy of two decimal places nine eigenvectors contain 100% of the variance for both the 2nd and 3rd quarters. The first seven include 98% and 97% of the variance. This is consistent with the interrelated definitions of the twelve targets.

Functional Models - General

A general functional form can be written as:

$y = f(x_1, x_2, \ldots x_n)$ Eq 1

The assertion that a functional form relates one metric to others is a relatively strong assertion. It requires that all the appropriate independent variables are included. The detection of fraud from an anomalous group of records assumes that the legitimate behaviour is determined by the function. If a number of possible 'targets' are considered, then one approach is to assume that there is a group of independent variables that determines all the targets in different functional forms. Another approach is to assert that the values must be internally consistent, with each having all others as the independent variables. In the first group one class of models corresponds to the independent measures not changing with any fraudulent behaviour, only the value of the target is different from the expected one. A more complex situation applies if the values of the input variables reflect fraudulent behaviour. The basis of this is a consistency between inputs and targets.

The functional models used in this study currently have some inputs, including the demographics, which are common to all functions. At least one of the other targets also appears as an independent variable in each function. Also the values of the inputs may change with fraudulent behaviour. For example one of the inputs is the number of patients. An extra patient will change this value. For this to be identified one or more of target variables need to be inconsistent with this new input value.

If a number of observations of all of the variables were taken then there are a number of techniques that can be used to obtain the function, or an estimate of the function. Some of these restrict the functional form. For others the functional form can only be easily accessed by predicting values from the independent variables. A technique that can theoretically fit any functional form is a sometimes known as 'universal function fitter (or learner)'. Two different examples of such learners are consider below; Feedforward Neural Networks and Boosted Regression Trees.

Testing strategy

The functional models used a four fold cross validation strategy with each subsample being a stratified one. The stratification was based upon the first principal component of the targets. The first stage was to partition the whole data set into three parts. Initially the number of services was used but this was changed to the total benefits. Of the three parts the lowest benefits are not considered for further analysis. This is related to the resources available for further examination as well as a lower financial cost of potential fraud in these cases. For each part the scores for the first principal component of the targets was calculated. This was used as input to a stratified random sampling routine and four partitions were obtained. Three were using in learning and the fourth was used as verification to indicate the anomalous behaviour.

If the behaviour is essentially unchanged over quarters it is possible to pass the new data through all four of the cross validation techniques and obtain an 'average' prediction. If there are only minor changes this can be done by saving the eigenvalues of the demographic group then basing the scores on the next quarter on these previous values. This procedure was not used as this data shows very different behaviour in the two quarters. The principal components, and the eigenvectors for the demographic groups are significantly different. As an example in the 3rd quarter some age ranges have a contribution for the most significant component with a negative value, in the 2nd quarter all have a positive value. The targets also differ between quarters. While the signs of the values of the eigenvector corresponding to the most important component are all positive in both cases the values differ by up to one third.

Feeedforward Neural Networks

Theoretically with enough data and enough nodes a single hidden layer feed forward network can learn any functional form. In practice multiple layers are used. These networks are essentially global function fitters. Neural networks are considered to be sensitive to extra or unnecessary input variables. The package used for neural networks was Neural Works Professional II Plus©. Both the inputs and outputs were scaled by the minimum and maximum. The transfer function for the internal nodes was an arctan. The outputs were defined as being linear. The default values of the learning parameters were used. Other networks and transfer functions were explored but none had any significant advantages over this choice. In practice neural networks learn best if the distribution of the examples over the range is relatively uniform. This was one reason why the tails of the distribution were considered separately.

It was decided that the targets would be considered as two groups, and a multiple output network used. Each group would have four networks constructed. A grouping has a disadvantage in that it is possible that one function would be influenced by another, crosstalk could be significant. Using a single network for each target would eliminate this possibility, but would require more networks and associated time. Another possible advantage from building multiple networks is that the inputs can be tailored separately for each output.

For the neural networks two of the random samples were used to train a variety of network architectures and the accuracy of these on other two were used to choose which architecture to use. The exploration of the number of hidden layers and the number of nodes used both the facilities of the software to remove extra nodes and human ex-

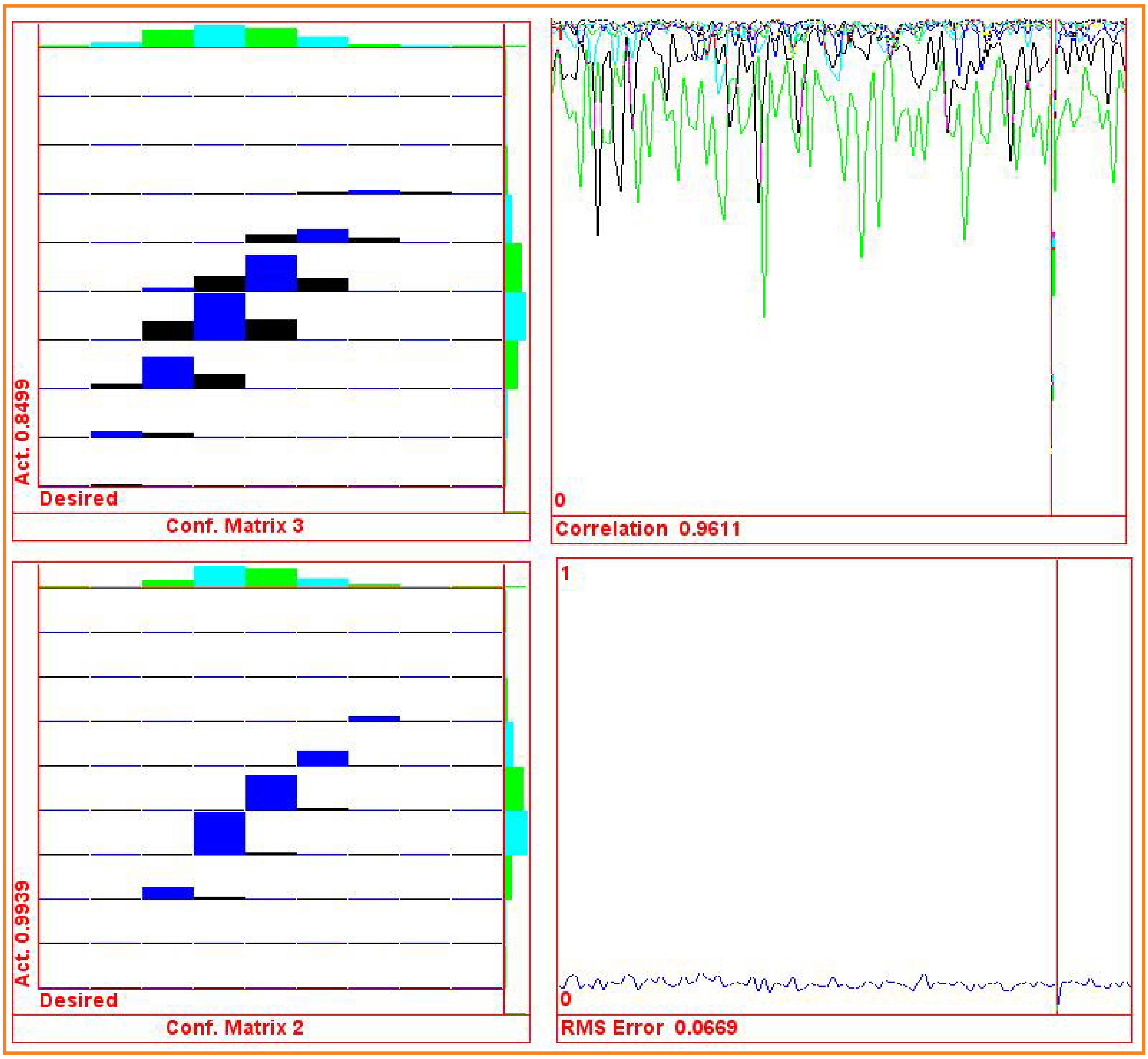

Figure 1: Selective screen shot of the accuracies using Neural Ware software

ploration of the number of layers, and nodes. The rms error, the correlation coefficients and the confusion matrices for each targets were used. The aims are to have a relatively low overall error, no single target to have a large error, and the historical value of the correlation coefficient should not have too many 'spikes'. Also the confusion matrices should be diagonally dominant (Figure 1). Another aim is to have a relatively simple network architecture. Once the architecture was chosen two random samples were used to train the network and he third was used to select the level of the training. Four networks were trained on the same architecture.

Boosted Regression Trees

Another technique for deriving a general function is through boosted regression trees. This technique is a local function fitter. Each tree projects the value of the residual as being constant within a leaf of the current tree. This leaf corresponds to a projection hypercube. The value of the function is thus constant in a small hypercube. Some implementation details of this version are described in Pearson 2000. Of particular relevance is the use of a test set to help in both error estimation and the decision to stop constructing extra trees. The proportion specified for the results described in this paper is one third for the test set. The stratified sampling within the software was used on a combined data set which corresponds to the three of the four stratified subsets. The last set used the trees saved from the construction to predict a value. Four sets of trees being used to obtain the predicted value for all examples.

Boosted regression trees can use the mean and the variance as a basis for choosing the best partition at a node. In this case the software allows a computationally efficient, but memory intensive option for a single initial sort. Initially all the values (N) of all variables (M) are sorted, a single pass of order $MNln(N)$. All successive nodes of the tree take of order N_n operations for each variable. For the more memory efficient computation each variable considered at node takes of order $N_n ln(N_n)$. In practice, because of the actual number of calculations this form is faster when there are small numbers at a node. This occurs in a single regression tree but not in the boosted form. A naive implementation takes N_n^2 operations at each node.

This form of calculation is not robust. A more robust version uses the median and the least absolute deviation (LAD). The current code takes a similar time for the calculation as a naive option for this form, of order $N_n^2 ln N_n$. More complex code, using linked lists for the sorted order and sorted order at each node is possible. This would still take of order N_n^2 for each variable. A compromise, included as an option, is to use the mean and variance for choosing the value within a variable and the median, LAD for the best variable.

Boosted regression trees are insensitive to extra variables. The only disadvantage is the time taken. Another feature of boosted regression trees is the ability to explicitly include a technique for missing values. When missing values occur these examples are not used in determining the best split at a node and a surrogate partition is used to partition the examples with missing values. In contrast Neural networks usually use a substitute value for both the training and estimate of the value of the function. The importance of the variables that are used in all the regression trees can be calculated. This can be used to select the variables to be used in the neural network.

Self Organising maps on the inputs

A Self Organising Map or Kohenen Neural network (usually denoted as a SOM) can be considered as a nonlinear fit of multiple dimensions onto two. In these two examples that are close in the original vector space are close in the Map. Thus SOM's were used as an aid in visualising the hyper-dimensional data Two examples that are close in the map are in a sense similar. Software also allows the coordinates from a SOM to be input to feed forward neural networks. For this data, two dimensions are insufficient for modelling purposes. The Map can however be used in helping with a visual presentation of the distribution of values and errors.

Two SOMs are calculated for each subset of the variables. Both have all of the standard input variables. One of these has benefits as an input and the other services.

Consequences of Function Fitting

Both techniques have are used to learn a function and then predict the values on the validation on all the data. The differences between the function values predicted and the actual can be conceptualised as consisting of two components. These are a modelling error and a learning error. Learning theory considers this latter form of error. It is sometimes called a probably approximately correct estimate. While a conceptual separation of the difference between the predicted and actual values is possible in practice only the total difference can be estimated.

The modelling error relates to the assumptions about the behaviour. Of particular interest are those examples where the modelling error relates to fraudulent behaviour. However this is not the only source of differences. There is often a valid reason why the theoretical model behaviour does not apply to a particular example.

In practice the learning error can be conceived as consisting of two aspects. One relates to the practical limitations of the functional form, the other relates to both the statistical noise in the data and the values of the parameters

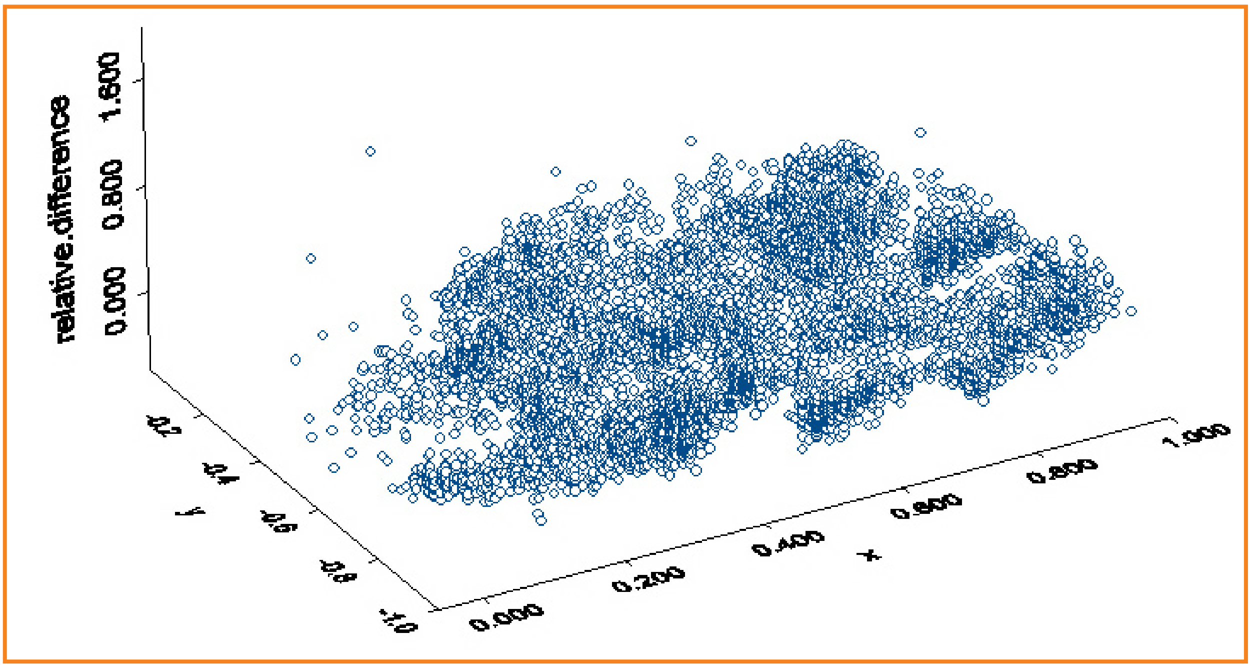

Figure 2: Relative difference of services expressed as a two-dimensional self-organising map.

in the representation of the function. The first of these is related to the limitations of the method in the choice of the function. For the neural networks the choice of the transfer function and the number of nodes and hidden layers both affect the function. For the Boosted regression trees each tree predicts a constant value in a (projection, limited dimension) hyper-rectangle. The final value can be considered as an additive model of these. Thus the number of boosted regression trees indicates how fine the constant parts can be. Another aspect of the learning error is the statistical error in evaluating the parameters of the limited functions. For Neural networks this is the approximate values that are found with the backpropogation (or other learning) technique. For boosted regression trees the errors relate both to the choice of the value, and the consequential standard errors.

In practice all conceptual parts of the errors co-exist. The differences related to inappropriate behaviour, (or model errors) the limitations of the function, and the approximation statistical errors cannot be separated.

The learning error is not necessarily uniform across the domain. Boosted regression trees are local learners and conceptually learn a constant value within a hypercube. In regions with many example points the hypercubes are small and where the learning data is sparse the regions are larger. The errors in learning a smooth function are smaller when the sample points for learning are close. It is important that the distribution of the points for learning are similar to those of new examples, while learning, and testing from a fractally distribution is more accurate than a uniform distribution (Pearson 2004). Boosted Regression Trees can also have lower errors than feedforward neural networks (Pearson 2000).

It is possible however to illustrate the differences in the predicted values of the two techniques. Firstly a SOM is used to project the inputs onto two dimensions. Then the relative differences between the predicted values of the two techniques is plotted against the coordinates. (figure) Multiple differences occur for each coordinate. In practice with this data the two coordinates are not sufficient for a functional form. In one of the ranges one of the values is clearly very different where the coordinate is separated from the others. This illustrates the differences between the local functional fitter with multiple constant values and a Global one with ˜smooth” continuous ones. The SOM technique was that provided by NeuralWorks Professional II Plus© software.

Difference Statistics

For each of the two chosen ranges of benefits, for each target variable and each of the two techniques estimates of the mean and variance of the differences between the predicted value of the function and the observed value were calculated. No firm pattern of the differences for each technique existed, the boosted regression tree usually had a smaller average error, and a larger variance than the neural network.

The absolute value of the difference in the mean can many be times larger for the neural network than the boosted regression trees. The target called services has errors of -8.04 and -25.32 for the ANN and -1.01 and -3.4 for the BRT. Similarly benefit with 347 and 570 compared to -13 and -7.

The variances of the boosted regression trees could be about 40% larger than the neural networks (e.g. the number of services in the medium benefits group had a variance estimate of 81.7 for the ANN and 117.94 for the BRT). There was only one of the 48 groups where the boosted regression trees had a greater absolute difference and that was for the high group on the number of items (-0.04 compared to -0.01 for the ANN). This same pair was a case where the ANN had the larger variance (1.56 compared to 1.66 for the ANN). The Diagnostic services for the medium benefits group is another example (10.69 compared to 11.48 for the ANN). The largest difference where the boosted regression trees were lower than the ANN was the average referred item for the medium group with 13.9 compared to 17.5.

These figures include all the contributions to the difference; The cases where the behaviour is atypical, the different functional forms, and the errors in estimating those forms.

Selection of Atypical Behaviour

After the mean and the standard deviation of the difference (m_i) has been calculated for each technique/group those examples where the value is more than two standard deviations from the mean are flagged as interesting. A measure of how atypical the example is also calculated. This measure (essentially a normalised L_1) is taken over all techniques and all variables. This is referred to as a suspicion score.

$$S = \sum_{i=1}^{24} \frac{|x_i^p - x_i^a|}{\mu_i} \qquad \text{Eq. 2}$$

The maximum value of this is found and a normalised measure of how much a particular example varies from the overall specified functional models is also calculated as

$$N_s = \frac{S}{\max(S)} \qquad \text{Eq. 3}$$

The suspicion score is also calculated for each technique while the normalised score is usually represented as a percentage.

The total number of times that the example has been flagged is also calculated.

Filtering

An atypical doctor may have a practice where the reason for the difference does not appear in the independent variables. Considerable resources are required for a detailed investigation of every case.

The number of flags per analysis technique are noted as are the total number of flags. The top 1000 or so providers are selected based upon the total number of flags, but mindful of the normalised average suspicion score (*see* equation 3).

The selection of providers to be investigated, is a two step process. In the first step the 1000 provider records were de-identified and randomly allocated to one of several Medicare Australia professional subject matter experts to look at. These are distributed as anomalous records. The records that were removed in the "top and tailing" of the distribution of the original data were similarly randomly distributed as outliers. The subject matter experts are required to indicate which providers are worth looking into further and why.

In step 2, those providers that were selected for follow-up, had their identifying details reassociated, and were reassigned to the subject matter experts responsible for the state or territory in which the primary practice was located. This expanded data set incorporated data from the providers practice over the preceeding year. This is included in recognition of the fact that a practitioners practice may be subject to major seasonal variations.

The state-based medical advisors evaluate the data and then must submit a case to the review group who may suggest a number of actions, including, but not limited to, doing nothing further, commissioning further investigations or sending a medical advisor to counsel the practitioner.

A simple database holds the providers summarised quarterly data, the number of flags set for each analysis technique, which variables they related to and their status at various stages of processing, including any comments attributed by the subject matter experts. Thus it is possible to express the relative effectiveness of each technique in isolation or when they agree upon the providers to look at. This approach also has the benefit of removing the possible bias towards alumni or colleagues, as the providers details will have been evaluated independently.

3 Results

The data outlined in tables 2-5 relate to the activity of general practitioners in the 3rd quarter of 2004. The analysis of the relative effectiveness of the techniques is currently incomplete as we are yet to complete the evaluation cycle. The results of the in depth analysis of the providers selected for further investigation is not yet known, but will act as feedback into our development cycle to enhance our predictive modelling accuracy.

From the original 28,496 providers, only 641 are worthy of following up (2.25 %), whilst the value rises to 3 percent if you consider only those providers actually working (i.e. services greater than 0). As you can see from Table 2, we are sending out to the state based medical advisors the details of approximately 16.8% of the providers that at least one of techniques has rated as being of interest because of the value of at least one variable. The subject matter experts indicated that almost 60% of the providers whose details we circulated in stage 1 of the processing were worth having a more in depth evaluation.

Of the 21337 active providers (delivered at least 1 service in the 3rd quarter of 2004), the artificial neural nets identified 3039 providers as being of interest (14.2%), whilst the boosted regression trees identified 5053 providers (12.3%). This is not indicative of the level of fraud within the medicare system, but indicates the numbers of providers that are considered to be unlike their peers. Often unusual patterns of practice highlighted by these techniques can be relatively easily explained with domain specific knowledge.

We would expect the number of flags per provider would be higher for those who the senior medical advisors have rated as being worthy of further investigation (Table 4).

This type of analysis may well only identify those who come to our attention because they have relatively high values for a number of variables.

The inclusion of the suspicion score was to highlight those providers who may not have many flags attributed by the ANN and BRT, but are close to the cut-off values for awarding flags over a number of variables (Table 5). It would therefore be expected that all measures of the suspicion score should be higher for those where the subject matter experts have agreed that the way the provider practices is worth further investigation.The analysis of those providers who were identified as falling into the outlier category resulted in 324 (71.8%) being rated as worthy of further analysis.

Table 2: Number of providers identified in each stage of the process, and the number remaining in the analysis. The data presented relates to the 3rd quarter of 2004 (July1 - September 30, 2004).

Stage of Processing	Number of Provider records	Records remaining in Analysis
Original extraction from Medicare	28496	28496
Inactive Providers (Services = 0)	7159	21337
Identify Outliers	149 (low) 451 (high)	20737
Lower Third (of data)	7331	13406
No Flags	6932	6474
Not Selected for Stage 1	5392	1082
No Follow-up	441	641

Table 3: Number of providers details that are present in the analysis as a function of the number of flags set for each analysis technique.

Records	ANN only	BRT only	Both	Total
	1421	2435	2618	6474
Stage 1 - No Followup	33	26	382	441
Stage 1 – Followup	27	23	591	641

Table 4: Result of Stage 1 analysis of anomalous providers in relation to the number of flags set per analysis technique and decision.

Decision	ANN Flags - average (range)	BRT Flags - average (range)
No Followup	2.4 (0 - 8)	2.1 (0 - 7)
Followup	2.8 (0 - 9)	2.5 (0 - 8)

4 Discussion

This system is one in a state of flux, taking note of the relative success of the different analysis techniques and their impact on the number of providers they have identified for further investigation. The analysis system outlined in this paper has not yet been operating for long enough to determine the number of providers who have been counselled or have been found to have defrauded the Medicare system. The strength of keeping meticulous records (via the database) is that we will be able to evaluate the relative accuracies of the various techniques in isolation and combined.

A number of the providers that are exonerated in the second stage of processing, may have been eliminated from further investigation because of the access to a years worth of practice data which may have smoothed out the peaks and troughs of the quarterly data.

Current techniques under investigation include the application of stochastic gradient boosting (Friedman, 2002). The inclusion of an expert system to sit between the evaluation techniques (ANN & BRT) and the circulation of the details of providers of interest (stage 1) is planed in an attempt to further reduce the incidence of known false positives.

We have already touched upon the future development of an overarching expert system that could attempt to classify the providers into certain categories. If a provider has a particular combination of flags than we could say something along the lines of if the provider is a type X provider then it is OK, otherwise there is a problem. Identification of sub-groupings within GP's could allow for such an approach. Inclusion of a case based reasoning engine would also allow more effective post-processing prior to circulation of individuals details to the subject matter experts. Over time the profile of providers and their practices could be extracted to provide a generalised profile of fraud within an area. Unfortunately, with the benefit of hindsight, we are now aware that once a certain loop hole is discovered and closed, another one will appear elsewhere, which someone will exploit for fraudulent purposes.

5 Conclusions

The impact of our fraud detection work has a flow-on effect amongst a provider's peers who may be deterred from continuing in their fraudulent behaviour, by the spotlight turned onto other providers undertaking similar behaviour.

This paper demonstrates the cumulative accuracy of applying two different techniques (a global and localised function fitter) to the data rather than

Decision	ANN Suspicion Score - average (range)	BRT Suspicion Score - average (range)	Average Suspicion Score -average (range)	Normalised Suspicion Score -average (range)
No Followup	16.5 (6.8 - 42.0)	14.5 (6.7 - 29.4)	15.5 (12.8 - 31.3)	33.8 (27.9 - 68.5)
Followup	17.6 (7.9 - 47.9)	15.7 (5.4 - 45.3)	16.7 (12.8 - 45.8)	36.4 (27.9 - 100)

Table 5: Result of Stage 1 analysis of anomalous providers in relation to their averaged normalised suspicion score per analysis technique and decision.

relying on one or the other technique in isolation.

Boosted Regression Trees and Neural Networks used together as function learners can be used to identify anomalous behaviour. Further analysis by experts is used to decrease the number that will be examined in more detail.

The following leasons have been learnt through the analysis undertaken during this project:

- The predictive power of the two techniques together, is better than if considered in isolation.
- The value of subject matter experts in being able to distinguish unusual behaviour from fraud.

This research emphasises the crucial importance of subject matter experts in real world applications. Approximately 40 percent of providers which the systems determined were worth a closer look, were believed to be explainable by the subject matter experts. In our system, the subject matter experts were senior medical advisors, all of whom who are qualified as general practitioners.

Acknowledgements

The authors wish to take this opportunity to acknowledge the contribution of other Medicare Australia staff, particularly Ms Edith Leung for retrieving the data from the electronic data warehouse and Mr Murray Cunningham for managing and massaging the data prior to the analysis. The authors also wish to acknowledge the critical role the Senior Medical Advisors of Medicare Australia continue to have on this on-going project.

References

Breiman, L. (2001). Random Forests. *Machine Learning*. 45(1): 5-32.

Copland, H. (2004). Regulatory applications of artificial intelligence. p 97-132 in Fulcher, J. and Jain, L.C. *Applied Intelligent Systems New Directions*. Springer-Verlag, Berlin.

Friedman, J.H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analytics*. 38: 367-378.

NAG (2002). NAG C Library Introductory Guide, Mark 7. The Numerical Algorithms Group. Oxford, U.K.

Pearson, RA. (2000). *Retrieval of Atmospheric Temperature Profiles from Satellite measurements Using Boosted Regression Trees*, p. 128-133 Proceedings of AISAT 2000, Karri, V. and Negevitsky M., (eds), Hobart, Australia.

Pearson, R.A. (2004). *Chaos, Fractals and Machine Learning*, Proceedings of CTAC/SIAM 2003, SIAM 2004, Sydney.

Semenova, T., Hegland, M. Graco, W. and Williams, G. (2004). *Conceptual mining of global large administrative health data*. p 659-669 in Dai, H., Srikant, R. and Zhang, C. PAKDD 2004. Springer-Verlag, Berlin.

Correspondence

Robert A. Pearson
Program Review Division,
Medicare Australia,
PO Box 1001
Tuggeranong DC ACT 2901

Robert.Pearson@medicareaustralia.gov.au