Predictive risk modelling for forecasting high-cost patients: a real-world application using Medicaid data

Sai T. Moturu*

Department of Computer Science and Engineering,
School of Computing and Informatics,
Arizona State University,
Tempe, AZ 85287-8809, USA
E-mail: smoturu@asu.edu
*Corresponding author

William G. Johnson

Center for Health Information and Research,
Department of Biomedical Informatics,
ASU Biomedicine,
Phoenix, AZ 85004-2430, Mail Code 8120, USA
E-mail: william.g.johnson@asu.edu

Huan Liu

Department of Computer Science and Engineering,
School of Computing and Informatics,
Arizona State University,
Tempe, AZ 85287-8809, USA
E-mail: huan.liu@asu.edu

Abstract: Approximately two–thirds of healthcare costs are accounted for by 10% of the patients. Identifying such high-cost patients early can help improve their health and reduce costs. Data from the Arizona Health Care Cost Containment System provides a unique opportunity to exploit state-of-the-art data analysis algorithms to mine data and provide actionable findings that can aid cost containment. A novel data mining approach is proposed for this challenging healthcare problem of predicting patients who are likely to be high-risk in the future. This study indicates that the proposed approach is highly effective and can benefit further research on cost containment.

Keywords: predictive risk modelling; healthcare expenditures; Medicaid; high-cost patients; data mining; non-random sampling; risk adjustment; skewed data; imbalanced data classification.

Reference to this paper should be made as follows: Moturu, S.T., Johnson, W.G. and Liu, H. (2010) 'Predictive risk modelling for forecasting high-cost patients: a real-world application using Medicaid data', *Int. J. Biomedical Engineering and Technology*, Vol. 3, Nos. 1/2, pp.114–132.

Biographical notes: Sai T. Moturu is a postdoctoral fellow in Computer Science and Engineering at Arizona State University. He has a PhD in Computer Science and Engineering and a Master's in Computational Biosciences from Arizona State University. His research is focused on the application of data mining and machine learning towards challenging problems in health informatics and social computing. He is currently working on the Social Health initiative at the Media Lab, Massachusetts Institute of Technology.

William G. Johnson is a Professor in the Department of Biomedical Informatics at Arizona State University. His research focuses on the development and application of community wide health information systems. He directs Arizona HealthQuery (AzHQ), a unique health data system that includes more than eight million citizens of Arizona. In addition, he directs projects including studies of the Arizona health workforce, a multi-state study of the causes and consequences of back pain and a study of healthcare disparities among Latino children.

Huan Liu is a Professor in the Department of Computer Science and Engineering at Arizona State University. He has extensive experience in research and development having worked in diverse countries like Singapore and Australia. His research focuses on data mining and machine learning with special emphasis on techniques such as feature selection, subspace clustering, ensemble methods and varied applications including recommender systems, bioinformatics, and social computing.

1 Introduction

The importance of electronic data has increased manifold over the last decade. With a consistent improvement in computing capabilities, substantial amounts of electronic data are being collected that can provide answers to critical research questions. In the high-impact field of healthcare informatics also, the utility of gathering electronic information has been realised. A part of this endeavour is a unique community health data system called Arizona HealthQuery (AzHQ), housed in the Centre for Health Information and Research (CHiR) at Arizona State University. AzHQ contains comprehensive health records of over eight million patients linked across time from the state of Arizona, USA. AZHQ offers the opportunity for research that can influence the community and deliver actionable results to researchers and policy makers.

The surge in data accumulation was accompanied by the development of improved methods of obtaining knowledge from the data. Data mining and machine learning methods have contributed immensely to this important task. These techniques have been successfully applied to many domains but applications to healthcare are still relatively rare (Anderson et al., 2004; Cios and Moore, 2002; Li et al., 2005). Using such techniques, this paper describes a potentially beneficial approach to healthcare, focusing on the difficult and important problem of predicting the patients who are likely to be high-risk.

One of the most important issues dogging the healthcare system is that approximately two–thirds of healthcare costs in the USA are accounted for by 10% of the patients (Berk and Monheit, 2001). This trend is not restricted to the USA alone and is common among all nations (Hammer, 1997). Aside from intractable problems, such as terminal cancer, the high-cost patients offer the greatest potential benefits in terms of cost reduction to strategies that include disease and disability management. Additionally, the costs of an episode of illness or injury among employed persons include losses of productivity due to work loss days and losses of on-the-job productivity. These 'indirect costs' are typically a multiple of healthcare costs for a given worker (Johnson, 2005).

The benefits are, however, unlikely to be realised unless the high-cost patients can be distinguished from lower cost patients early in the process of care. Proactive identification of high-cost patients can help design targeted interventions and disease management programs suited to the high-risk population in question. Such health promotion programs are known to reduce health risks for many patients, resulting in improved patient health as well as cost reduction from both direct and indirect costs (Musich et al., 2000). Predictive risk modelling techniques can help forecast such high-cost patients.

2 Related work

The leading approaches to risk modelling include Adjusted Clinical Groups (ACG), Diagnostic Cost Groups (DCG), Global Risk-Adjustment Model (GRAM), RxRisk, and Prior Expense. The utility of these models arises from their designs as well as their use of predictors. These models, which are designed to use domain knowledge and expertise, yield comparable results. To predict whether a patient is high-risk or not, these models use healthcare utilisation information and disease-related features or morbidity indicators based on diagnoses codes and other administrative claims-based data (Meenan et al., 2003).

Demographic variables like age and sex are known to impact healthcare costs. Disease-related predictors from various utilisation classes such as inpatient, outpatient and pharmacy have also been used to predict cost outcomes. It was found that using data from multiple utilisation classes provides better predictions (Zhao et al., 2005). Patient's health conditions can also be used as predictors in the form of comorbidity indices. The predictive ability of various comorbidity indices was found to be similar (Perkins et al., 2004). It was also found that simple count measures like number of prescriptions and number of claims were better predictors of healthcare costs than comorbidity indices (Farley et al., 2006).

Many studies rely on regression methods rather than risk adjustment models for prediction. Regression techniques generally tend to predict the average cost for a group of patients satisfactorily but on an individual basis; the predictions are not too accurate because healthcare cost distributions do not meet the assumptions of normality and homoscedasticity. Cost distributions are frequently transformed to avoid these problems (Diehr et al., 1999). Despite these issues, a comparison of multiple models for the estimation of future total healthcare costs using pharmacy claims data found that ordinary least squares regression was a better approach when compared to more complex models (Powers et al., 2005).

Unlike existing approaches, this study uses data mining techniques for predictive modelling that learn from data to tailor suitable models. Data mining has been successfully applied previously in many applications including financial applications like fraud detection in credit card transactions, stock market prediction, portfolio management, bankruptcy prediction, and identifying trading rules in the foreign exchange market (Zhang and Zhou, 2004). Despite the success, data mining has scarcely been applied to healthcare informatics. However, it is particularly suited to this problem because imbalanced data are commonly observed in many applications like credit card fraud detection, network intrusion detection, insurance risk management, text classification, and medical diagnosis that have been widely studied by the data mining and machine learning community (Chawla et al., 2004). The problem with such data is that most classification algorithms assume that class distribution is uniform. Particularly, the metric of classification accuracy is based on this assumption. This means that algorithms often try to improve this faulty metric while learning. Therefore, it is essential to pay attention to unbalanced class distributions when dealing with claims-based healthcare expenditure data.

The two most common solutions to this problem include non-random sampling (under-sampling or down-sampling, over-sampling or up-sampling and a combination of both) and cost-sensitive learning. Both solutions have a few drawbacks (most importantly, under-sampling might neglect some key instances while over-sampling might result in overfitting) but they were equally successful in showing improvement over conventional techniques (McCarthy et al., 2005; Weiss and Provost, 2001). The use of synthetic examples for the minority class was found to show improvement over under-sampling or over-sampling in certain cases (Chawla et al., 2002) but it is not prudent to use this technique for healthcare data where highly varied instances tend to group together.

Comparisons of over-sampling, under-sampling and cost-sensitive learning are inconclusive. Some find that there is little difference in the results from these data sets (Maloof, 2003), others usually find that one of them is better (Batista et al., 2004; Drummond and Holte, 2003; McCarthy et al., 2005). While it is hard to pick an option due to these contrasting results, it was also suggested that the use of a combination of under-sampling and over-sampling that balances training data shows improved performance (Estabrooks et al., 2004). In addition, it was found using varying ratios of the minority and majority classes for a data set that the best results were generally obtained when the minority class was over-represented in the sample (Weiss and Provost, 2001). We explore the possibility of using a combination of over-sampling and under-sampling together with classification algorithms for the creation of predictive models that can forecast high-cost patients. Preliminary work has confirmed the usefulness of this approach (Moturu et al., 2007, 2008).

3 Our approach

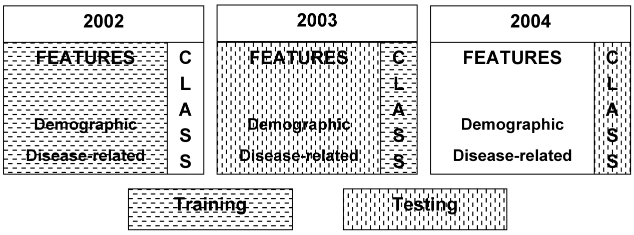
Our approach for risk modelling consists of three major parts that mirror the steps in a typical knowledge discovery task using data mining. The initial step is data preprocessing, which is considered by some to be the most important step in the process. This step entails data selection and preparation for the creation of suitable test and training data. The Arizona Health Care Cost Containment System (AHCCCS), Arizona's Medicaid program was selected for this study as it satisfies the requirement for a multi-year administrative claims-based data set containing disease-related information from various utilisation classes for demographically varied patients.

A de-identified sample from the AHCCCS data for three years (2002–2004) including 437 features from 139,039 persons was extracted from AzHQ for this study. Demographic features include age category (ages in groups of five), gender, county, race and marital status. Disease-related features from four utilisation classes including inpatient, outpatient, emergency department and pharmacy were used. For pharmacy data, the National Drug Code (NDC) classification was used to group information into 136 categories. For the other utilisation classes, procedure codes from the International Classification of Diseases (ICD) were used to group information into 20 Major Diagnostic Categories (MDC). For each of these classes, information is available as the number of visits. This was used to create two sets of variables, one set that include actual visit counts and another set that include a binary value indicating the presence or absence of visits in the category. We refer to these variables as visit counts and binary indicators respectively.

The patients were categorised into the minority or rare class (high-cost) and the majority class based on the amounts paid for healthcare. The practice of discounting billed charges in the healthcare industry requires that the amounts paid for the services are used as measures of costs rather than the amounts charged. Two thresholds of \$50,000 (0.69% high-cost patients) and \$25,000 (2.18% high-cost patients) were selected to ensure sufficiently skewed data.

Figure 1 depicts the structure of the data and its division. As the goal is to predict whether a patient would be high-risk in the following year using information from the current year, features from one year and class from the following year were used together for both learning and evaluation. Training data are constructed with features from 2002 and class from 2003 while test data are constructed with features from 2003 and class from 2004.

Figure 1 Illustration of training and test data



The skewed nature as well as the large size of the data necessitates sampling as a part of the data preparation process. Accordingly, non-random sampling was used as a combination of over-sampling and under-sampling to create both balanced and imbalanced training data. Random sampling was also used for comparison. Further details are provided in Section 5.

Following the creation of suitable training and data, the next step in the process is model learning. Classification techniques can be used to learn models that can classify patients into the majority and minority classes. Based on preliminary testing using a variety of such techniques including Bayesian and decision tree methods, only five were found to perform well. These algorithms include AdaBoost (with 250 iterations of a Decision Stump classifier), LogitBoost (also with 250 iterations of a Decision Stump classifier), Logistic Regression, Logistic Model Trees, and the Support Vector Machine

(SVM) classifier. These techniques were used to learn predictive models using the training data.

The final step in the process involves testing and evaluation of these predictive models. The unabridged test data described above is used to test these models and evaluate their predictions. Performance evaluation provides a considerable challenge for predictive modelling since traditional measures of success like accuracy are not suitable in this case. We describe these evaluation techniques in detail in the following section.

4 Evaluation

The following four evaluation metrics are proposed to gauge the performance of our predictive models quantitatively:

- *Sensitivity*. Sensitivity corresponds to the proportion of correctly predicted instances of the minority class with respect to all such instances of that class. It is equal to the number of true positives over the sum of true positives and false negatives.

$$S_T = \frac{N_{TP}}{N_{TP} + N_{FN}}.$$

- *Specificity*. Specificity corresponds to the proportion of correctly predicted instances of the majority class with respect to all such instances of that class. It is equal to the number of true negatives over the sum of true negative and false positives.

$$S_P = \frac{N_{TN}}{N_{TN} + N_{FP}}.$$

- *F-measure*. *F*-measure is typically used as a single performance measure that combines precision and recall and is defined as the harmonic mean of the two. Here we use it as a combination of sensitivity and specificity.

$$F_M = \frac{2 \times S_T \times S_P}{S_T + S_P}.$$

- *G-mean*. *G*-mean typically refers to geometric mean. As the *F*-measure described above, it is a single performance measure that is used to combine specificity and sensitivity using geometric mean instead of the harmonic mean.

$$G_M = \sqrt{S_T \times S_P}.$$

To evaluate the performance of these models, it is necessary to understand the relevance of their predictions. The predictions from these models are used to reallocate resources such that the high-risk patients are carefully looked after with specially designed case management and intervention programs. The intent is to take special care of the high-risk patients without neglecting low-risk patients in the process resulting in healthier patients that ensure a reduction in both direct and indirect costs. Consider the following example of two models created using non-random and random sampling. Table 1 depicts the predictions from these models. In the first scenario, the model developed using a non-random sample correctly identifies 675 high-cost patients (70.8% sensitivity) while incorrectly predicting 21,812 patients as high-cost (84.2% specificity). In the second scenario, the model developed using a random sample correctly identifies 32 high-cost patients (3.4% sensitivity) while incorrectly predicting 82 patients as high-cost (99.9% specificity). In the second scenario, 32 patients might benefit upon resource reallocation but the remaining 96.6% high-cost patients are unidentified. Hence, large portions of the health and cost benefits are unattainable. In contrast, there is a strong possibility that many more patients are benefited in the first scenario.

Table 1 Random vs. non-random sampling: an example

	Non-random sample		*Random sample*	
	Minority	*Majority*	*Minority*	*Majority*
Predicted minority	*675*	21,812	*32*	82
Predicted majority	279	*116,273*	922	*138,003*

This example depicts the need for high sensitivity. However, a drop in specificity generally accompanies increased sensitivity. This drop also needs to be monitored to ensure that too many low-risk patients are not predicted to be high-risk. This brings about the need for an acceptable trade-off between specificity and sensitivity as can be evaluated by the *F*-measure or *G*-mean. However, this is only a guideline and it would be hard to identify the best trade-off. Hence, we also provide a way to obtain variable trade-offs from which an acceptable one can be picked when background knowledge about a suitable trade-off is available.

The four measures described above evaluate the performance of predictive models based on number of correct and incorrect predictions. Since costs are the focus of this study, performance of these models needs to be assessed further in terms of costs. We propose to compare these models using the proportion of costs captured correctly by their predictions. This would be equal to the sum of costs from the correctly predicted patients in a class over the total sum of costs from the patients in that class. Such a measure was previously used with risk adjustment models to assess the quality of predictions for the high-risk class (Meenan et al., 2003). An increase in the quality of predictions for the high-risk class is generally accompanied by a decrease in the same for the low-risk class, much like sensitivity and specificity. Hence, both these values need to be evaluated simultaneously. Therefore, we extend a similar measure to the low-risk class also. We use C_H and C_L to refer to these proportions for the high and low risk classes respectively.

$$C_H = \frac{\sum_{\text{correctly predicted}} \text{Cost (minority class)}}{\sum_{\text{all patients}} \text{Cost (minority class)}}$$

$$C_L = \frac{\sum_{\text{correctly predicted}} \text{Cost (majority class)}}{\sum_{\text{all patients}} \text{Cost (majority class)}}.$$

Apart from quantitative evaluation measures, we devise a qualitative evaluation method to identify the distance between actual cost and predicted class using cost categories. We define eight cost categories, four for each class to identify how far away the wrong predictions are. For each class, we calculate the proportion of patients that are incorrectly classified. This proportion is equal to the total number of incorrect predictions in that category over the total number of patients observed in that category. Ideally, the percentage of wrong predictions in a cost category should be smaller if the distance of that category from the threshold is larger. Such an evaluation can be visualised using line charts that are capable of picturing these trends.

5 Empirical study: design and results

A data mining approach to predictive risk modelling was proposed based on preliminary results. This approach includes the use of non-random sampling followed by model learning and evaluation as described in the previous sections. These elements set the stage for an empirical study that can provide further insight into this approach by testing its credibility while also depicting its flexibility and consistency. All experiments in the study were performed using the Weka software (Witten and Frank, 2005). This study can be used to answer the following specific questions with respect to the prediction of high-cost patients:

- How important is the contribution of non-random sampling to this predictive modelling approach? Can the performance improvement due to non-random sampling be observed using different classification algorithms?
- Can the process of non-random sampling be modified to gain further improvement in performance? Is there a suitable trade-off between sensitivity and specificity that can be achieved by the models learned using this approach?
- How important is the information from each utilisation class and which class is most useful?
- How important is the individual visit information to the overall performance?
- Which is the most suitable classification algorithm for this approach? Is its performance consistent?
- Is this approach robust to changes in the threshold used to separate the high-cost patients?

5.1 Importance of non-random sampling

The first group of experiments were designed to depict the importance of non-random sampling, as it is the key to the success of this approach. To provide a foundation for comparison, a base line method for predictive modelling was devised wherein the future cost outcome of a patient is assumed to be the same as the current outcome. Any method that is to be deemed useful would require better performance than this base line that has a sensitivity of 0.276 and specificity of 0.993 ($F_M = 0.432$, $G_M = 0.524$). For the creation of non-randomly sampled training data, 20 random samples of the data were obtained from both classes separately. Each of these samples contained a 1000 data points resulting in a training data set with 40,000 data points. A set of experiments were also performed using random sampling in place of non-random sampling to indicate the improvement in performance. In this case, half of the data was randomly selected to create a training data set. Disease-related information was represented using binary indicators.

Table 2 depicts the results from this comparison. Random sampling results in poor performance with a sensitivity that is much lower than the base line rendering this method useless. On the other hand, non-random sampling shows a large improvement in sensitivity over the base line clearly indicating its usefulness. These observations are complemented by the *F*-measure and *G*-mean.

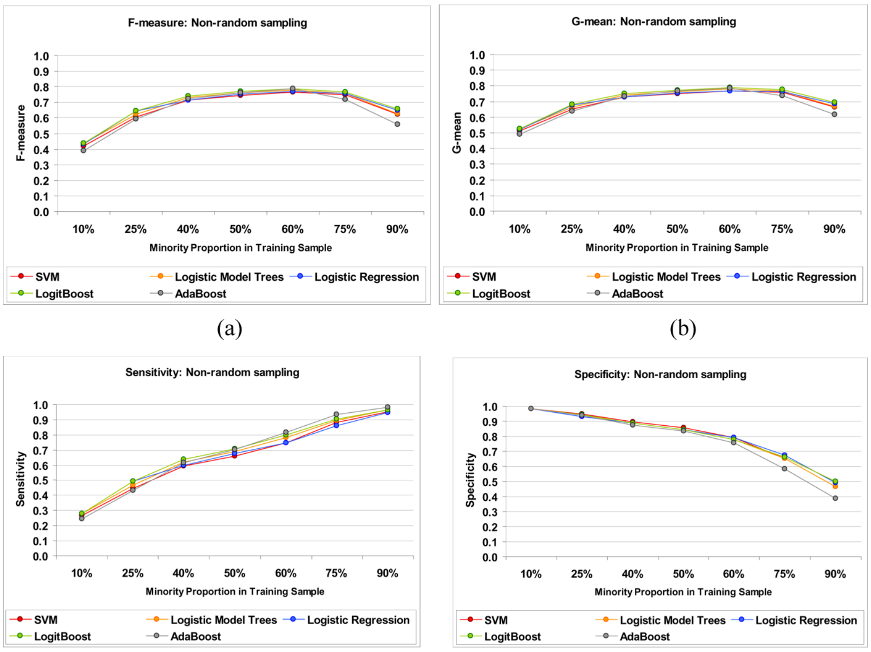
Table 2 Performance comparison among training samples

			Non-random (*Minority:Majority*)					
		Random	*10:90*	*25:75*	*40:60*	*50:50*	*60:40*	*75:25*	*90:10*
AdaBoost	*ST*	*0.021*	*0.244*	0.433	0.614	*0.701*	0.818	*0.933*	*0.982*
	SP	1.000	0.983	0.938	0.876	0.835	0.756	0.583	0.389
	F_M	0.041	0.391	0.592	0.722	0.762	0.786	0.718	0.557
	GM	0.145	0.490	0.637	0.733	0.765	0.786	0.738	0.618
	CH	*0.033*	*0.271*	0.466	0.635	*0.726*	0.826	*0.939*	*0.984*
	C_L	0.997	0.920	0.782	0.637	0.559	0.434	0.234	0.106
LogitBoost	*ST*	*0.034*	*0.281*	0.492	0.636	*0.708*	0.90	*0.906*	*0.964*
	SP	0.999	0.981	0.942	0.886	0.842	0.779	0.662	0.501
	FM	0.065	0.437	0.6	0.741	0.769	0.789	0.765	0.659
	GM	0.183	0.525	0.7510	0.751	0.772	0.789	0.775	0.695
	CH	*0.059*	*0.309*	0.522	0.671	*0.734*	0.817	*0.915*	*0.967*
	C_L	0.996	0.910	0.784	0.4620	0.567	0.462	0.316	0.177
Logistic regression	*ST*	*0.035*	*0.279*	0.492	0.950	*0.677*	0.748	*0.862*	*0.950*
	SP	0.999	0.981	0.930	0.889	0.845	0.791	0.675	0.493
	FM	0.067	0.434	0.643	0.716	0.752	0.769	0.757	0.649
	GM	0.186	0.523	0.76	0.7630	0.756	0.769	0.763	0.685
	CH	*0.061*	*0.305*	0.517	0.621	*0.718*	0.87	*0.877*	*0.955*
	C_L	0.995	0.910	0.750	0.662	0.4950	0.495	0.342	0.190
Logistic model trees	*ST*	*0.000*	*0.280*	0.465	0.621	*0.692*	0.780	*0.897*	*0.965*
	SP	1.000	0.981	0.941	0.887	0.844	0.779	0.653	0.463
	F_M	0.000	0.436	0.623	0.760	0.760	0.779	0.756	0.626
	GM	0.000	0.524	0.662	0.742	0.764	0.779	0.765	0.669
	CH	*0.000*	*0.304*	0.492	0.635	*0.711*	0.796	*0.903*	*0.967*
	C_L	1.000	0.911	0.785	0.654	0.576	0.304	0.304	0.164
SVM	*ST*	*0.002*	*0.267*	0.448	0.592	*0.657*	0.746	*0.883*	*0.953*
	SP	1.000	0.984	0.947	0.896	0.857	0.791	0.651	0.464
	$\mathrm{F_M}$	0.004	0.420	0.608	0.713	0.744	0.768	0.749	0.624
	GM	0.046	0.513	0.651	0.729	0.750	0.769	0.758	0.665
	CH	*0.003*	*0.295*	0.466	0.631	*0.679*	0.768	*0.891*	*0.955*
	C_L	1.000	0.921	0.800	0.679	0.605	0.494	0.313	0.177

Balancing the training data to tackle data imbalance is an intuitive idea that has worked well in the past and with our data. However, it was suggested in the past that increasing the representation of the minority class in the training sample might improve performance. We test this notion by varying the class distributions of the minority an d majority classes in the training sample. The total training sample was kept consistent at 40,000 data points with individual random samples of thousand each. However, the number of these samples for each class was varied based on the proportions of rare class being considered for a particular experiment. For example, if the rare class proportion was 25%, ten samples from the minority class were used resulting in 10,000 data points with the other 30 samples belonging to the majority class. Six different non-randomly sampled training data sets were created using rare class proportions of 10%, 25%, 40%, 60%, 75% and 90% in addition to the existing balanced training sample.

Table 2 lists the detailed results from this comparison. It can be observed that using a higher proportion of minority class data in the training sample does provide certain improvement in performance as depicted by the *F*-measure and *G*-mean in Figure 2(a) and (b). Our results indicate that a moderate increase in proportion of the minority class (between 50% and 75%) improves performance in terms of these measures but using extreme proportions of the minority class (90%) are not suited to this application. This is in slight contrast to the observations of Weiss and Provost (2001) who observe that class distributions with the minority class percentage as high as 90% are useful with their data.

Figure 2 Performance comparison for various training samples: (a) *F*-measure; (b) *G*-mean; (c) sensitivity; (d) specificity; (e) proportion of costs predicted correctly in the minority class and (f) proportion of costs predicted correctly in the majority class (see online version for colours)



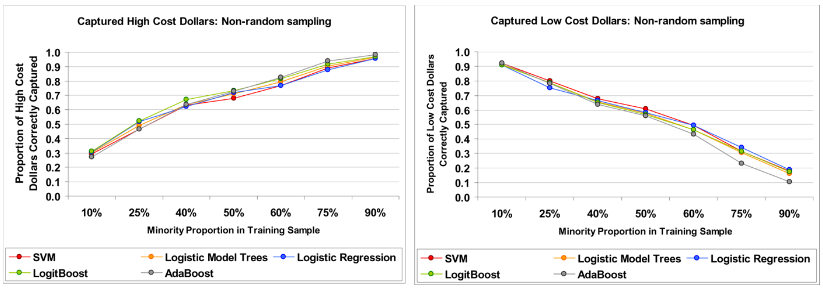
(a)

(b)

(c)

(d)

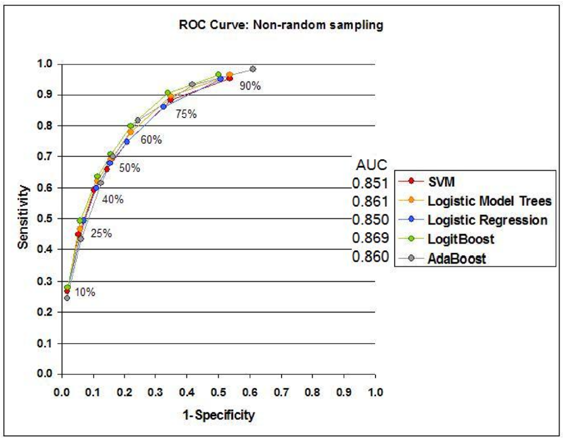
Figure 2 Performance comparison for various training samples: (a) *F*-measure; (b) *G*-mean; (c) sensitivity; (d) specificity; (e) proportion of costs predicted correctly in the minority class and (f) proportion of costs predicted correctly in the majority class (see online version for colours) (continued)



(e) (f)

Figure 2(c) and (d) show the trends in sensitivity and specificity as the proportion of minority class is increased in the training sample. Sensitivity is consistently improved with such an increase while specificity deteriorates. As discussed earlier, one needs to find the best trade-off between these measures. Figure 3 depicts a Receiver Operating Characteristics (ROC) curve that aptly visualises the trade-off between sensitivity and specificity. With such a curve, the point closest to the top left corner is considered the best. In this case, this point is the class distribution with 60% minority class instances in the training data. The observations from this curve coincide with our previous inference that a moderate imbalance in the training data in favour of the minority class can improve performance.

Figure 3 ROC curve (see online version for colours)

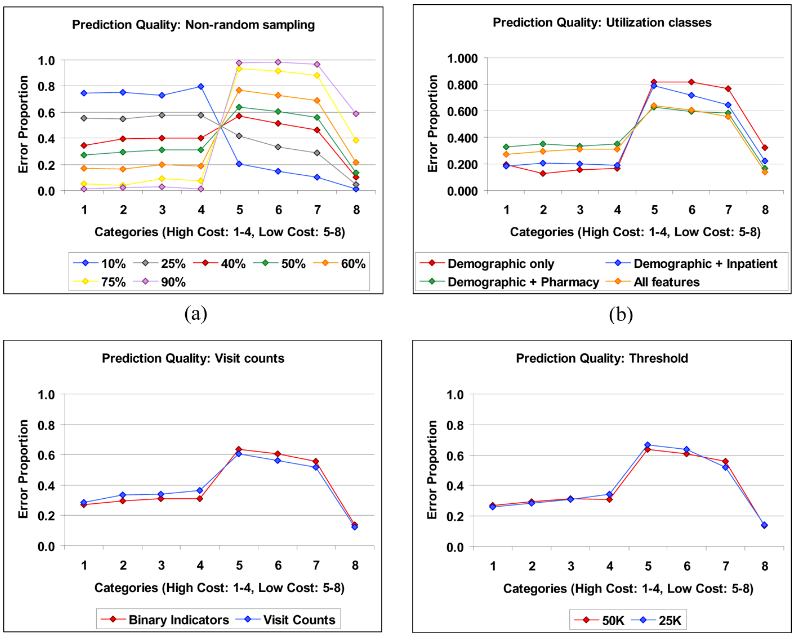


As described earlier in Section 4, there is a requirement for high sensitivity for this application but such sensitivity is not acceptable if there is a considerable drop in specificity. Though the *F*-measure, *G*-mean and the ROC curve point toward the usefulness of changing the class distribution in the training sample, this may not always be the case. These measures give equal importance to specificity and sensitivity but different real world situations could demand different trade-offs between these measures.

It is here that the approach of customising the class distributions in the training sample can have immense impact. With the availability of real world data that can suggest a suitable trade-off for a particular situation, one can change these distributions as necessary to obtain the appropriate trade-off. This flexibility makes our approach a very useful tool that can be used by practitioners to create models tailored to meet the requirements for varied situations.

Further, Figure 2(e) and (f) depict the trends in the percentages of cost correctly predicted for both classes. It is interesting to note that these curves closely follow the trends observed for sensitivity and specificity. A large percentage of the costs from high-risk patients are predicted correctly. However, it is important to assess the quality of these predictions. Using the measure described in Section 4, costs in both classes are divided into four categories with the expectation that categories farthest to the threshold would show the least percentage of error. This is depicted in Figure 4(a). The error percentages for minority class predictions are close in the four categories with a slight downward trend away from the threshold. It can be observed that the experiments with a larger percentage of minority class training instances show a consistent decrease in error percentages. The exact opposite trend is observed for majority class percentages. A large percentage of the prediction errors are observed in the categories closer to the threshold with a downward trend for categories further away. This measure helps confirm the quality of predictions from models created by our approach. It can be observed that there is a higher error percentage for majority class predictions closer to the threshold. This is not necessarily bad as these patients could potentially be high-risk and these categories have a lower number of patients as well. However, this indicates that there is scope for improvement in the quality of predictions for the majority class.

Figure 4 Prediction quality for various experiments: (a) non-random sampling; (b) utilisation classes; (c) visit counts and (d) threshold (see online version for colours)



(a)

(b)

(c)

(d)

5.2 Significance of utilisation classes

Data from different utilisation classes were used in the past to predict cost outcomes, especially from inpatient and pharmacy data. It was also suggested that using data from multiple utilisation classes could provide improved performance. In this group of experiments, we test the significance of these utilisation classes individually as well as using combinations among them. Four sets of experiments were performed, each of them utilising demographic information. Balanced training data was created for each of these experiments using non-random sampling. Disease-related information was represented using binary indicators. The first set of experiments included no disease-related information. The next set of experiments included inpatient information alone followed by another set that included pharmacy information alone. The final set of experiments included disease-related information from all the four utilisation classes.

Table 3 depicts the results from these comparisons. As expected, the use of only demographic information is the least useful as depicted by the *F*-measure and *G*-mean. A high sensitivity is achieved coupled with a low specificity. Nevertheless, this result is striking because it manages such numbers despite the use of little information. This is particularly promising because it provides a basic method to help categorise patients when prior disease-related information is unavailable. The relative usefulness of pharmacy and inpatient information with our approach is detected by the experiments that use such information along with demographic data. It was found that inpatient information results in a better performance than pharmacy data in terms of the *F*-measure and *G*-mean as well as in terms of quality as depicted in Figure 4(b). Surprisingly, the use of only inpatient information also provides a slight improvement in performance in terms of *F*-measure and *G*-mean over the use of all disease-related information for experiments using three of the five classification algorithms. These results indicate the usefulness of inpatient information with our approach but also suggest that there is scope for pharmacy information to be better utilised. With respect to the use of disease information from multiple utilisation classes, one can conclude that this provides a performance that is very close to the best, if not the best. If data from different utilisation classes is available, it is suggested that all such information is used for the creation of predictive models. However, our results show that even without the presence of data from multiple utilisation classes, useful predictive models can be created. These results reiterate the flexibility of this approach for predictive modelling, which implies that it could be used for varied data sets with differing disease-related information.

Table 3 Performance comparison among utilisation classes

		Demographic only	*Demographic + inpatient*	*Demographic + pharmacy*	*All features*
AdaBoost	*ST*	0.836	0.806	0.660	0.701
	SP	0.646	0.750	0.808	0.835
	FM	*0.729*	*0.777*	*0.727*	*0.762*
	GM	*0.735*	*0.778*	*0.730*	*0.765*
	CH	0.835	0.812	0.681	0.726
	CL	0.343	0.459	0.536	0.559

Table 3 Performance comparison among utilisation classes (continued)

		Demographic only	*Demographic + inpatient*	*Demographic + pharmacy*	*All features*
LogitBoost	*St*	0.831	0.816	0.668	0.708
	Sp	0.645	0.747	0.829	0.842
	Fm	*0.726*	*0.780*	*0.740*	*0.769*
	Gm	*0.732*	*0.780*	*0.744*	*0.772*
	Ch	0.831	0.815	0.680	0.734
	Cl	0.343	0.455	0.566	0.567
Logistic regression	*St*	0.763	0.731	0.622	0.677
	Sp	0.661	0.766	0.824	0.845
	Fm	*0.708*	*0.748*	*0.709*	*0.752*
	Gm	*0.710*	*0.748*	*0.716*	*0.756*
	Ch	0.750	0.733	0.653	0.718
	Cl	0.364	0.483	0.562	0.580
Logistic Model trees	*St*	0.767	0.731	0.631	0.692
	Sp	0.658	0.766	0.832	0.844
	Fm	*0.708*	*0.748*	*0.718*	*0.760*
	Gm	*0.710*	*0.748*	*0.724*	*0.764*
	Ch	0.756	0.733	0.640	0.711
	Cl	0.360	0.482	0.572	0.576
SVM	*St*	0.777	0.764	0.600	0.657
	Sp	0.659	0.750	0.845	0.857
	Fm	*0.713*	*0.757*	*0.702*	*0.744*
	Gm	*0.715*	*0.757*	*0.712*	*0.750*
	Ch	0.758	0.765	0.622	0.679
	Cl	0.359	0.460	0.600	0.605

5.3 Influence of visit counts

Intuition suggests that the availability of visit counts would prove useful for predictive modelling. We test this perception using two sets of experiments. Balanced training data was created for each of these experiments using non-random sampling. The first set of experiments use visit counts for each disease-related feature while the second set of experiments use binary indicators for the same feature as explained in Section 3. This comparison is intended to test the usefulness of visit counts.

It can be observed from Table 4 that using binary indicators provides a slightly lower specificity combined with a higher sensitivity compared to the use of visit counts. Though one may expect that count measures should provide better results, both the *F*-measure and *G*-mean are marginally higher with the use of binary indicators. This unexpected result could imply that specific visit counts are not as important as one would expect. However, the understanding that required trade-offs between sensitivity and specificity might not give equal importance to both these measures means that visit counts could provide better performance in certain situations. An analysis of the prediction quality, depicted in Figure 4(c), shows that both sets of experiments show similar error proportions in the various categories. Hence, one can only infer that the knowledge of presence or absence of disease in the various categories is enough to create good predictive models when visit counts are unavailable. This makes our approach more flexible in terms of the features used.

Table 4 Performance comparison between visit counts and binary indicators

	AdaBoost		*LogitBoost*		*Logistic regression*		*Logistic model trees*		*SVM*	
	Binary	*Count*	*Binary*	*Count*	*Binary*	*Count*	*Binary*	*Count*	*Binary*	*Count*
S_T	*0.701*	0.668	*0.708*	0.646	*0.677*	0.646	*0.692*	0.632	*0.657*	0.594
S_P	*0.835*	0.850	*0.842*	0.894	*0.845*	0.899	*0.844*	0.902	*0.857*	0.919
F_M	*0.762*	0.748	*0.769*	0.750	*0.752*	0.752	*0.760*	0.743	*0.744*	0.722
G_M	*0.765*	0.753	*0.772*	0.760	*0.756*	0.762	*0.764*	0.755	*0.750*	0.739
C_H	*0.726*	0.701	*0.734*	0.675	*0.718*	0.683	*0.711*	0.671	*0.679*	0.633
C_L	*0.559*	0.593	*0.567*	0.663	*0.580*	0.667	*0.576*	0.669	*0.605*	0.711

5.4 *Effect of class thresholds*

Despite the success of our approach with a highly skewed data set, one could question the robustness of this approach to differently skewed data. To test this view, we use a second class threshold, as described in Section 3. We perform two sets of experiments using these two thresholds for comparative analysis. Balanced training data was created for each of these experiments using non-random sampling. Disease-related information was represented using binary indicators.

It can be observed from Table 5 that the results from both thresholds are comparable with the greater threshold ($50,000) providing slightly better performance. Though one might expect that the lower threshold provides better performance due to the decrease in imbalance, it is not the case here because with this approach the training data are balanced for both situations. This indicates that the slight decrease in performance from the data set with a lower threshold could be attributed to the fact that there are more patients around this threshold, increasing the chance for prediction error. Figure 4(d) shows the prediction quality for both cases. It is very interesting to note that both curves follow each other closely despite that fact that the cost categories used are different due to the obvious difference in threshold. This clearly depicts the robustness of our approach to changes in class threshold indicating the flexibility of our approach towards differently skewed data.

Table 5 Performance comparison between thresholds

	AdaBoost		*LogitBoost*		*Logistic regression*		*Logistic model trees*		*SVM*	
	50K	*25K*	*50K*	*25K*	*50K*	*25K*	*50K*	*25K*	*50K*	*25K*
S_T	*0.701*	0.681	*0.708*	0.696	*0.677*	0.666	*0.692*	0.678	*0.657*	0.647
S_P	*0.835*	0.809	*0.842*	0.819	*0.845*	0.833	*0.844*	0.820	*0.857*	0.837

Table 5 Performance comparison between thresholds (continued)

	AdaBoost		*LogitBoost*		*Logistic regression*		*Logistic model trees*		*SVM*	
	50K	*25K*	*50K*	*25K*	*50K*	*25K*	*50K*	*25K*	*50K*	*25K*
F_M	*0.762*	0.740	*0.769*	0.753	*0.752*	0.741	*0.760*	0.743	*0.744*	0.730
G_M	*0.765*	0.742	*0.772*	0.755	*0.756*	0.745	*0.764*	0.746	*0.750*	0.736
C_H	*0.726*	0.708	*0.734*	0.727	*0.718*	0.701	*0.711*	0.707	*0.679*	0.682
C_L	*0.559*	0.542	*0.567*	0.548	*0.580*	0.574	*0.576*	0.553	*0.605*	0.582

5.5 Performance of classification algorithms

Though non-random sampling is a key step in our approach, model learning using classification algorithms assumes equal importance in the process of predictive modelling. Based on preliminary analyses, five algorithms were found to perform similarly in conjunction with non-random sampling. These five algorithms were used through all our experiments to test their consistency as well as to identify the best of these. Throughout the various comparisons, the performance of these algorithms is very similar despite the various changes in features as well as quantitative and qualitative evaluation measures. Such an observation strengthens our selection of these algorithms. However, this makes it difficult to pick the best among these. Figure 3 provides a comparison among these algorithms using area under the ROC curve (AUC) values. Using this metric, LogitBoost provides the best performance, followed by Logistic Model Trees and AdaBoost. Nevertheless, any of these five algorithms can be used as a part of our predictive modelling approach to achieve good results consistently.

6 Conclusions and future work

Predictive risk modelling for forecasting high-cost patients is an important area of research and this study provides a useful data mining approach for the task. A comprehensive empirical study of this new technique using a real-world Medicaid data set tested by multiple qualitative and quantitative evaluation measures provides ample confirmation of its merits. Results indicate that the use of non-random sampling to create training data helps balance the challenges resulting from the skewed nature of healthcare expenditure data sets. Further, changing the class distribution in the training sample to moderately increase representation of the minority class helps improve performance further. Though this study manifests the significance of non-random sampling in building predictive risk models, it is difficult to select the best model as the identification of the most suitable trade-off between specificity and sensitivity is problematic without specific data on the cost benefits to be gained from such models. The technique of using varied class distributions in the training samples to create alternative models as a part of our approach is a blessing because it provides the capability to adapt to varied real world situations that could require diverse trade-offs between these measures. When data on cost benefits is available, one can test various class distributions to select a suitable one that can aid in the creation of a predictive model with the best cost benefit. This makes our approach for predictive modelling much more adaptable.

Our approach can create models automatically by learning from the data and is therefore not restricted to the use of a specific type of data or features. This approach can make use of as much (data from multiple utilisation classes) or as little data (data from a single utilisation class or data without visit counts) as available and still prove useful. A comparison of classification algorithms for this task indicates that the selected five work almost equally well. Though LogitBoost seems to provide the best performance, the other algorithms are not far behind. Therefore, our results indicate to future users a handful of appropriate classification techniques to be used along with non-random sampling for predictive modelling. Further, the threshold for high-cost patients is tuneable and can be varied depending on the goals of a particular study. Hence, this approach can be easily adapted to diverse studies with different features and varied levels of imbalance in the class distributions. All these taken together signify the flexibility and consistency of our customisable approach for predictive risk modelling and the benefits that can be obtained from such analyses.

Considering the variation in data, predictors and evaluation metrics, comparison with previous studies is improper. Nevertheless, our results are better (double the sensitivity at about the same level of specificity) than a decision-tree based predictive modelling technique (Anderson et al., 2004). The ROC curve in Figure 3 is similar to the one obtained for existing risk-adjustment models (Meenan et al., 2003). The best AUC value for that study is 0.86, which is equalled by our approach while using AdaBoost and bettered while using Logistic Model Trees and LogitBoost. Additionally, the best risk-adjustment models are found to predict about 30% of the high-cost dollars correctly, when the threshold for high-cost patients is 0.5%. Our models, on the other hand, correctly predict over 70% of the high-cost dollars with a balanced training sample when the threshold for the high-cost patients is 0.69%. These comparisons validate the usefulness of this technique that is further enhanced by its adaptability. As can be observed, non-random sampling is the most important component of this technique and is very beneficial for the creation of suitable predictive models.

Predictive risk modelling is a useful technique with practical application for numerous employers and insurers in the goal to contain costs. We provide a promising approach that is valuable, flexible and proven to be successful on real-world data. Nevertheless, there is further scope to improve the interpretation of these results. It is commonly observed that a considerable percentage of high-cost patients do not remain that way every year. In addition, two patients could share very similar profiles with only one of them being high-cost. Studying these seemingly anomalous patients could provide a better understanding of how a high-cost patient is different from other patients. Further, feature selection can be considered in order to eliminate features that do not significantly affect the outcome. Feature weighting is another viable option that is more intuitive for this application as different features affect the outcome to different extents. In addition, the current sampling approach and available classification techniques could be further refined to enhance performance.

The most promising possibilities for the extension of this work arise from working closely with key data partners. This avenue provides the opportunity to incorporate information on the cost containment methods used and their efficiency as well as real data on the cost benefits obtained from previous predictive models. The use of expert knowledge through interaction with these partners can help tailor this method further to suit varied real world requirements. Working with such partners, we endeavour to provide a reasonable, patient-specific answer to this question that would significantly influence cost containment in the healthcare industry.

Acknowledgements

We thank the members of the Data Mining and Machine Learning Lab and the Centre for Health Information and Research at Arizona State University for their constructive criticism and helpful comments regarding this work. Sai T. Moturu and Huan Liu's work are partially sponsored by AFOSR and NSF.

References

Anderson, R.T., Balkrishnan, R. and Camacho, F. (2004) 'Risk classification of Medicare HMO enrollee cost levels using a decision-tree approach', *Am. J. Managed Care*, Vol. 10, No. 2, pp.89–98.

Batista, G.E.A.P.A., Prati, R.C. and Monard, M.C. (2004) 'A study of the behavior of several methods for balancing machine learning training data', *ACM SIGKDD Explorations Newsletter*, Vol. 6, No. 1, pp.20–29.

Berk, M.L. and Monheit, A.C. (2001) 'The concentration of health care expenditures, revisited', *Health Affairs*, Vol. 20, No. 2, pp.9–18.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, Vol. 16, pp.321–357.

Chawla, N.V., Japkowicz, N. and Kolcz, A. (2004) 'Editorial: special issue on learning from imbalanced data sets', *ACM SIGKDD Explorations Newsletter*, Vol. 6, No. 1, pp.1–6.

Cios, K.J. and Moore, G.W. (2002) 'Uniqueness of medical data mining', *Artificial Intelligence in Medicine*, Vol. 26, Nos. 1–2, pp.1–24.

Diehr, P., Yanez, D., Ash, A., Hornbrook, M. and Lin, D.Y. (1999) 'Methods for analysing health care utilization and costs', *Ann. Rev. Public Health*, Vol. 20, pp.125–144.

Drummond, C. and Holte, R.C. (2003) 'C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling', *ICML Workshop on Learning from Imbalanced Datasets II*, Washington DC, USA.

Estabrooks, A., Jo, T. and Japkowicz, N. (2004) 'A multiple resampling method for learning from imbalanced data sets', *Computational Intelligence*, Vol. 20, No. 1, pp.18–36.

Farley, J.F., Harley, C.R. and Devine, J.W. (2006) 'A comparison of comorbidity measurements to predict healthcare expenditures', *Am. J. Manag. Care*, Vol. 12, pp.110–117.

Hammer, J.S. (1997) 'Economic analysis for health projects', *The World Bank Research Observer*, Vol. 12, No. 1, pp.47–71.

Johnson, W.G. (2005) 'Cost-based evaluations of the treatment of back pain: a primer for health care professionals', *The Spine Journal*, Vol. 5, No. 4, pp.361–369.

Li, J., Fu, A.W., He, H., Chen, J., Jin, H., McAullay, D., Williams, G., Sparks, R. and Kelman, C. (2005) 'Mining risk patterns in medical data', *Proc 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD'05)*, pp.770–775.

Maloof, M. (2003) 'Learning when data sets are imbalanced and when costs are unequal and unknown', *ICML Workshop on Learning From Imbalanced Datasets II*, Washington DC, USA.

McCarthy, K., Zabar, B. and Weiss, G. (2005) 'Does cost-sensitive learning beat sampling for classifying rare classes?', *Proc. 1st Int'l Workshop on Utility-based Data Mining (UBDM '05)*, pp.69–77.

Meenan, R.T., Goodman, M.J., Fishman, P.A., Hornbrook, M.C., O'Keeffe-Rosetti, M.C. and Bachman, D.J. (2003) 'Using risk-adjustment models to identify high-cost risks', *Med. Care*, Vol. 41, No. 11, pp.1301–1312.

Moturu, S.T., Johnson, W.G. and Liu, H. (2007) 'Predicting future high-cost patients: a real-world risk modeling application', *Proc. IEEE International Conference on Bioinformatics and Biomedicine 2007*, pp.202–208.

Moturu, S.T., Liu, H. and Johnson, W.G. (2008) 'Healthcare risk modeling for medicaid patients: the impact of sampling on the prediction of high-cost patients', *HEALTHINF 2008*, January, pp.126–133.

Musich, S.A., Adams, L. and Edington, D.W. (2000) 'Effectiveness of health promotion programs in moderating medical costs in the USA', *Health Promotion International*, Vol. 15, No. 1, pp.5–15.

Perkins, A.J., Kroenke, K., Unutzer, J., Katon, W., Williams Jr., J.W., Hope, C. and Callahan C.M. (2004) 'Common comorbidity scales were similar in their ability to predict health care costs and mortality', *J. Clin. Epidemiology*, Vol. 57, pp.1040–1048.

Powers, C.A., Meyer, C.M., Roebuck, M.C. and Vaziri, B. (2005) 'Predictive modeling of total healthcare costs using pharmacy claims data', *Med. Care*, Vol. 43, No. 11, pp.1065–1072.

Weiss, G.M. and Provost, F. (2001) *The Effect of Class Distribution on Classifier Learning: An Empirical Study*, Tech Report ML-TR-44, Dept. Computer Science, August, Rutgers University.

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, San Francisco.

Zhang, D. and Zhou, L. (2004) 'Discovering golden nuggets: data mining in financial application', *IEEE Trans. Sys. Man Cybernet. C Appl. Rev.*, November, pp.513–522, Vol. 34, No. 4.

Zhao, Y., Ash, A.S., Ellis, R.P., Ayanian, J.Z., Pope, G.C., Bowen, B. and Weyuker, L. (2005) 'Predicting pharmacy costs and other medical costs using diagnoses and drug claims', *Med. Care*, January, pp.34–43.