Predicting Future High-Cost Patients: A Real-World Risk Modeling Application

Sai T. Moturu[*#], William G. Johnson[#+], Huan Liu[*]
[*]*Department of Computer Science and Engineering*
[#]*Center for Health Information & Research (CHiR)*
[+]*Department of Biomedical Informatics*
School of Computing and Informatics, Arizona State University
Tempe, AZ 85287-8809, USA
smoturu@asu.edu, william.g.johnson@asu.edu, hliu@asu.edu

Abstract

Health care data from patients in the Arizona Health Care Cost Containment System, Arizona's Medicaid program, provides a unique opportunity to exploit state-of-the-art data processing and analysis algorithms to mine the data and provide actionable results that can aid cost containment. This work addresses specific challenges in this real-life health care application to build predictive risk models for forecasting future high-cost users. Such predictive risk modeling has received attention in recent years with statistical techniques being the backbone of proposed methods. We survey the literature and propose a novel data mining approach customized for this potent application. Our empirical study indicates that this approach is useful and can benefit further research on cost containment in the health care industry.

1. Introduction

The Center for Health Information and Research (CHiR) at Arizona State University houses a community health data system called Arizona HealthQuery (AZHQ). AZHQ contains comprehensive health records of patients from the state of Arizona linked across systems and time. The data, which include more than six million persons, offer the opportunity for research that can impact on the health of the community by delivering actionable results for health care researchers and policy makers.

The different players that shape the health care market include employers, insurers, care providers, patients and suppliers. While containing costs is of high priority to insurers and purchasers, health care providers and suppliers would prefer to resist such cost containment. Since a large percentage of the patients depend on employers for health insurance, employers and insurers are primarily responsible for the incurred costs. Apart from health care costs, employers also incur other health-related expenditures due to the time taken by the patient to return to work and the resulting loss of productivity. These "indirect costs" provide further incentive for employers to care for the health of their employees and the quality of health care provided.

The probable reasons for the consistent growth in health care expenditures range from the lack of a free market and the development of innovative technologies to external factors like economy and population growth [1]. However, most of these factors are often difficult to alter and a relatively easier goal is to devise effective cost containment measures. It is well known that a high proportion of health care costs are associated with a small proportion of patients. This phenomenon has been observed for health related indirect costs in employed populations. One efficient approach to contain costs is to focus on high-cost patients responsible for these expenditures and undertake measures to reduce costs. Predictive risk modeling is a relatively recent attempt at identifying high-cost patients to contain costs. We embark on the challenging task of building (learning) predictive risk models using real-life data from the Arizona Health Care Cost Containment System (AHCCCS), Arizona's Medicaid program, available in AZHQ.

The AHCCCS data was selected because it contains a large number of patients that can be tracked over multiple years and it contains the features required for this study. Apart from the challenge for data analysis

due to the voluminous amount of patient records and the considerable amount of variation among similar patients, such cost data provides a bigger challenge due to the skewed cost distribution. It has been observed that the top ten percent of the population accounts for more than two-thirds of the health expenditures and this has remained consistent over the years [2].

Since a tiny percentage of patients create a large portion of the impact, identifying them beforehand would allow for the design of better cost containment measures. Note that historical data provides better information about a patient than predictive models but neither is used for discriminatory purposes. Early identification using predictive models can help design targeted interventions and more effective disease and case management programs for high-risk patients that can help defer or even avoid adverse outcomes. From the perspective of the employers, better return on investment could be achieved due to the reduction in indirect costs while insurers are helped by the reduction in risk. Such predictions could also help establish capitation reimbursements. However, the data imbalance provides a challenge for such prediction.

As a part of this study, we propose a predictive risk modeling approach to identify high-risk patients and use the results as a basis for developing predictions of the probability that an individual would become a high-cost patient. We borrow from the field of data mining and machine learning to design such an approach as it has been successfully applied in many applications including financial applications [3]. However, it has not been regularly used in the field of health care data analysis. We study the possibility of applying some of these data mining techniques to aid in predictive risk modeling, where we aim to predict the outcome of a patient for the next year based on data from the current year. We propose an approach for the same using a combination of sampling and classification techniques.

2. Related Work

2.1 Learning from imbalanced data

The primary problem while analyzing health care data with respect to expenditures is the highly skewed nature of the data. The problems of dealing with imbalanced data for classification have been widely studied by the data mining and machine learning community [4]. Most classification algorithms assume that the class distribution in the data is uniform. Particularly, the metric of classification accuracy is based on this assumption and algorithms often try to improve this faulty metric while learning from imbalanced data. It is essential to pay attention to this fact while dealing with health care 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 are equally successful over conventional techniques [5, 6].

Different studies have compared the usefulness of over-sampling, under-sampling and cost-sensitive learning in the past. While some suggest that there is little difference in their outcomes, others indicate that one of them is better. These contrasting results have made it difficult to pick one of these as a better option. In addition, the combination of under-sampling and over-sampling is also found to be useful [6, 7, 8, 9, 10]. Despite the success of these techniques in other domains, none of them have been applied to health care expenditure data in the past. In this study, we explore the possibility of using non-random sampling as a key element while learning predictive risk models.

2.2 Current techniques

Health care data sets have been used in the past for the prediction of future health care utilization where the goal varied from predicting individual expenditures to estimating total health care expenditures. Typically, various regression techniques have been employed for this task with varying success. Ordinary least squares regression is used in many studies but its assumptions are not satisfied by the skewed distribution of the costs. These techniques generally tend to predict the average cost for a group of patients satisfactorily but on an individual basis, the predictions are not very accurate. Other approaches include transforming the distribution to match the assumptions of the analysis technique and using the Cox proportional hazards model [11].

Risk-adjustment models that can forecast individual annual health care expenses are also available. These can predict high-cost patients using a cost threshold. The utility of these models arises from the predictors employed. They use utilization data and disease-related features or morbidity indicators based on diagnoses codes and other claims-based data [12]. Though current techniques for the analysis of health care cost data are predominantly statistical, data mining could prove just as useful and we explore this possibility.

2.3 Predictors of high-cost

Health care data sets from different sources have been used to predict future utilization. Self-reported health status data gathered from surveys has been used to predict expenditures [13] and group patients into cost categories [14]. Instead, administrative claims data is employed in this study. Demographic variables like age and sex are known to work well as predictors for expenditure. Apart from such predictors, disease-related information from various utilization classes such as inpatient, outpatient and pharmacy has been used earlier, either separately or together to predict cost outcomes. It has been suggested that using multiple utilization classes provides better predictions [15]. Comorbidity indices that quantify the diseases or conditions possessed by the patient have also been used as predictors. However, recent studies suggest that the use of multiple utilization classes and the use of simple count measures like number of claims or prescriptions were found to be better predictors of health care costs [16, 17]. Though the performance of such indices may vary, disease information is still a key predictor.

3. Predictive Risk Modeling

3.1 Data and features

The size of AZHQ necessitates the selection of a specific subset for analysis. The primary requirement is for a multi-year administrative claims-based data set containing disease-related information from various utilization classes. It is also essential that there is demographic variation among patients. Based on these requirements, AHCCCS data that provides a large sample size of 139039 patients was selected. A total of 437 demographic and disease-related features were extracted from the original AHCCCS data for three consecutive years (2002 to 2004). All the features in the data set were either categorical or binary. The patients were categorized into the high-cost and lower cost classes based on the paid amount. Since the goal is to predict future costs, features from one year and class from the following year have been used together. Training data was constructed with features from 2002 and class from 2003 while test data was constructed with features from 2003 and class from 2004.

Demographic variables including age category (ages in groups of five), gender, county, race and marital status have been used. Age and gender have been included based on previous success. Race, location and marital status have been included due to their possible impact on both health and financial aspects. Disease-related data from various utilization classes including inpatient, outpatient, emergency department and pharmacy is used. For inpatient, outpatient and emergency department data, procedure codes from the International Classification of Diseases (ICD) have been grouped into twenty major diagnostic categories (MDC). For the pharmacy data, the classification has been derived from the National Drug Code (NDC) classification with 136 categories. For each of these 196 features, information is available as a count of the number of visits or as a binary value to indicate if there have been any visits within the category. We refer to these as visit counts and binary indicators respectively.

Since our interest is in predicting high-cost patients, it is necessary to separate patients into classes. The practice of discounting billed charges in the health care industry requires that the amounts paid for services are used rather than the amounts charged. Hence, payments are used in this study. To differentiate cost classes, we use thresholds of $50000 (0.69% or 954 high-cost patients) and $25000 (2.18% or 3028 high-cost patients) that ensure highly skewed data.

3.2 Challenges for predictive modeling

There are three major challenges for building predictive models. The first is from data imbalance that invariably results in poor performance with conventional analysis techniques. The selection of appropriate classification techniques provides the second challenge. The unbalanced nature of the data also brings about the final challenge - the selection of suitable evaluation metrics to gauge the performance of the models created by these algorithms.

To address the challenge provided by data imbalance, non-random sampling has been employed to create a balanced training sample. A combination of over-sampling the minority class (high-cost patients) and under-sampling the majority class (lower costing patients) is employed to create a balanced sample with an equal number of patients from both classes. This approach is reasonable as it has proven successful in the past. The large data size is also tackled by sampling. Training data is thus sampled from the data with features from 2002 and class from 2003.

The next challenge is model learning. We have tested a variety of popular classification algorithms to focus on the challenge of learning from the training data. Out of these algorithms, five worked considerably better. These include the Support Vector Machine (SVM) classifier, Logistic Regression, Logistic Model Trees, AdaBoost and LogitBoost (the last two used a Decision Stump classifier and 250 iterations).

The models learned from training data using these algorithms are used to predict on the test data with features from 2003 and class from 2004. The performance evaluation of these models provides the

final challenge. Traditional measures of success like accuracy are not useful as the data is highly skewed. We propose to use the following evaluation metrics:

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

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

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

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

3. 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. In this study we use a single performance measure that combines sensitivity and specificity and refer to it as F-measure.

$$F_M = \frac{2 * S_T * S_P}{S_T + S_P}$$

To evaluate the performance of these models, the relevance of their predictions needs to be understood. 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 21812 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). Using these models, employers would reallocate their resources such that the high-risk patients are carefully looked after with specially designed case management and intervention programs. In the second scenario, 32 patients might benefit but the remaining 96.6% high-cost patients are unidentified. Hence, a large portion 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. This example depicts the need for an acceptable tradeoff between specificity and sensitivity as can be evaluated by the F-measure, with a specific need for high sensitivity.

To summarize, we outline our three-pronged approach to predictive risk modeling for clarity. The first preprocessing step includes the use of non-random sampling to create balanced training data. In the second model building step selected classification algorithms are used to learn models from the training data. The final test step involves model testing that is evaluated by the three selected measures. This sets the stage for an empirical study on predictive risk modeling. In the following section, we first describe our experimental design and then discuss experimental results.

Table 1: Confusion Matrix - Random vs. Non-random Sampling

	Non-Random Sample		Random Sample	
	Positive	Negative	Positive	Negative
Predicted Positive	**675**	21812	**32**	82
Predicted Negative	279	**116273**	922	**138003**

4. Empirical Study

4.1 Experimental design

Employing the AHCCCS data, an extensive range of experiments are conducted using subsets of the data to provide a comprehensive comparative outlook. All experiments have been performed using the Weka software [18]. The data set with a threshold of $50,000 for class separation containing all available binary-valued disease features was used for the default set of experiments. Non-random sampling was used to create training data as a default. For each experiment, all selected classification algorithms have been used to create predictive models to identify the best one. The following dimensions were used for comparison.

4.1.1 Sampling techniques. Experiments were designed to depict a clear difference in performance of the sampling techniques. One set of experiments were performed using random sampling where 50% of the data was randomly selected for learning. Another set of experiments were performed using non-random sampling where 20 random samples were obtained for both classes, with every sample containing 1000 data points. The resulting data contained 40000 data points.

4.1.2 Combinations of information from different utilization classes. To assess the best combination of information from various utilization classes, four sets of experiments were performed. These experiments contained a varying amount of data including experiments using demographic information only, demographic and inpatient information, demographic and pharmacy information, and the default set with all the pieces of available information.

4.1.3 Visit counts and binary indicators. Separate data sets with either the visit counts or binary indicators for each disease-related feature were used for these experiments to gauge the difference between the two. Results are provided for the default experiment.

4.1.4 Thresholds for class differentiation. Two different thresholds ($50000 and $25000) were used to differentiate cost classes to assess whether the technique is robust to variations along this boundary. Results are provided for the default experiment.

4.2. Results and discussion

4.2.1. The importance of non-random sampling. Both random and non-random samples are drawn from the same data to form training data in order to build predictive models. The purpose of this experiment is to verify the usefulness of non-random sampling. Table 2 provides the results from this comparison. It is apparent that random sampling provides very poor sensitivity with less than 10% of the high-cost patients being identified correctly. Additionally, we consider a baseline model where patients are predicted to be in the same class as they were in the previous year. Such a model provides a sensitivity of 0.276 and a specificity of 0.993 for this data (an F-measure of 0.432). Low sensitivity indicates that not many high-cost patients remain in that category for the next year making predictive modeling more difficult. The sensitivity for the baseline model is much better than that achieved from random sampling but poorer than that from non-random sampling. Non-random sampling helps in achieving sensitivity as high as 0.7 but as one would expect, this comes with a loss in specificity. However, the F-measure is much higher (over 0.7 for all five algorithms) indicating that the tradeoff between sensitivity and specificity is the best among the three options. These results clearly depict the effectiveness of non-random sampling for predictive modeling.

Table 2: Random vs. Non-random Sampling

		S_T	Sp	F_M
AdaBoost	Random	0.021	1	0.041
	Non-Random	**0.701**	**0.835**	**0.762**
LogitBoost	Random	0.034	0.999	0.065
	Non-Random	**0.708**	**0.842**	**0.769**
Logistic Regression	Random	0.035	0.999	0.067
	Non-Random	**0.677**	**0.845**	**0.752**
Logistic Model Trees	Random	0	1	0
	Non-Random	**0.692**	**0.844**	**0.76**
SVM	Random	0.002	1	0.004
	Non-Random	**0.657**	**0.857**	**0.744**

4.2.2 Selected classification algorithms perform well. Five classification algorithms were used to learn predictive models for all the experiments with the purpose of identifying the best among them. Recall that these algorithms were selected over many others from our preliminary analysis. Tables 2, 3, 4 and 5 depict that these five techniques perform consistently well across all dimensions. Though the F-measure indicates that results from the LogitBoost algorithm seem marginally better, one can only conclude that any of these five could be used to learn a suitable predictive model from non-randomly sampled training data.

Table 3: Using different combinations of disease-related features

		S_T	Sp	F_M
AdaBoost	Demographic	0.836	0.646	0.729
	Demo + Inpatient	**0.806**	**0.75**	**0.777**
	Demo + Pharmacy	0.66	0.808	0.727
	All	**0.701**	**0.835**	**0.762**
LogitBoost	Demographic	0.831	0.645	0.726
	Demo + Inpatient	**0.816**	**0.747**	**0.78**
	Demo + Pharmacy	0.668	0.829	0.74
	All	**0.708**	**0.842**	**0.769**
Logistic Regression	Demographic	0.763	0.661	0.708
	Demo + Inpatient	**0.731**	**0.766**	**0.748**
	Demo + Pharmacy	0.622	0.824	0.709
	All	**0.677**	**0.845**	**0.752**
Logistic Model Trees	Demographic	0.767	0.658	0.708
	Demo + Inpatient	**0.731**	**0.766**	**0.748**
	Demo + Pharmacy	0.631	0.832	0.718
	All	**0.692**	**0.844**	**0.76**
SVM	Demographic	0.777	0.659	0.713
	Demo + Inpatient	**0.764**	**0.75**	**0.757**
	Demo + Pharmacy	0.6	0.845	0.702
	All	**0.657**	**0.857**	**0.744**

4.2.3. Using different combinations of disease-related information. Multiple combinations of disease-related information from different utilization classes are compared in Table 3. The purpose of this experiment is threefold: (1) to gauge the utility of demographic information (2) to identify whether inpatient or pharmacy information is more useful (3) to identify the combination with the best performance. From among these combinations, the use of only demographic information is the least useful as depicted by the F-measure (0.726 while using the LogitBoost algorithm for the default experiment). This provides a high sensitivity (0.831) accompanied by a significantly lower specificity (0.645). Nevertheless, this result is striking because it manages such high numbers despite the use of little information and has much better sensitivity than the baseline model and random

sampling using all information. This is particularly promising because it provides a way to categorize patients when prior information is unavailable. Adding pharmacy data results in a comparable F-measure (0.74). However, this time the sensitivity (0.668) is lowered while the specificity is higher (0.829). Using inpatient data instead of pharmacy data causes a drop in specificity (0.747) and a significant improvement in sensitivity (0.816). Surprisingly, the F-measure (0.78) in this case is the best among the four comparisons. The use of all disease-related and demographic information results in a slightly lower F-measure (0.769) while achieving greater specificity (0.842) and lower sensitivity (0.708). Despite the varying tradeoffs, these combinations provide promising results indicating the flexibility of our approach, implying that predictive models can be created from varied data sets with differing disease-related information.

4.2.4 Binary-valued disease-related features are useful. Disease-related features using both binary indicators and visit counts were compared. Visit information is not always available and this comparison helps inspect the performance in such cases. It can be observed from Table 4 that using binary indicators provides a slightly lower specificity (0.842 to 0.894 in the case of the LogitBoost algorithm) and a higher sensitivity (0.708 to 0.646 for the same case) compared to the use of visit counts. Though one may expect that count measures should provide much better results, our experiments show that the F-measure is marginally higher while using binary indicators (0.769 to 0.750). This unexpected result is consistent across the five algorithms and could imply that visit counts are not as important as one would expect. Instead, binary indicators can result in good predictive models.

4.2.5 Robustness to changes in class threshold. Two different thresholds were used for class separation to indicate the robustness of this technique to changes in this threshold. This comparison is interesting because we observe from Table 5 that the results for the higher threshold consistently show better sensitivity and specificity (0.708 and 0.842 as opposed to 0.696 and 0.819 while using the LogitBoost algorithm). One would expect that the lower threshold improves the balance between classes providing better results. This is true in the case of random sampling where the sensitivity (0.069 to 0.034) and F-measure (0.129 to 0.065 while using the LogitBoost algorithm) show a marked improvement in the data set with the lower threshold. However, with non-random sampling, the training data is already balanced and hence, it does not affect the results. Since there are more patients closer to the lower threshold, it is likely that there is a higher chance of error. This could be the reason for the slight underperformance while using non-random sampling with a less skewed data set. This comparison indicates that the threshold can be adapted as required with varied data sets while still achieving similar results.

Table 4: Disease-related features: Binary indicators vs. Visit counts

		S_T	Sp	F_M
AdaBoost	Counts	0.668	0.85	0.748
	Binary	**0.701**	**0.835**	**0.762**
LogitBoost	Counts	0.646	0.894	0.75
	Binary	**0.708**	**0.842**	**0.769**
Logistic Regression	Counts	0.646	0.899	0.752
	Binary	**0.677**	**0.845**	**0.752**
Logistic Model Trees	Counts	0.632	0.902	0.743
	Binary	**0.692**	**0.844**	**0.76**
SVM	Counts	0.594	0.919	0.722
	Binary	**0.657**	**0.857**	**0.744**

Table 5: Performance of different thresholds

		S_T	Sp	F_M
AdaBoost	Threshold: $25000	0.681	0.809	0.74
	Threshold: $50000	**0.701**	**0.835**	**0.762**
LogitBoost	Threshold: $25000	0.696	0.819	0.753
	Threshold: $50000	**0.708**	**0.842**	**0.769**
Logistic Regression	Threshold: $25000	0.666	0.833	0.741
	Threshold: $50000	**0.677**	**0.845**	**0.752**
Logistic Model Trees	Threshold: $25000	0.678	0.82	0.743
	Threshold: $50000	**0.692**	**0.844**	**0.76**
SVM	Threshold: $25000	0.647	0.837	0.73
	Threshold: $50000	**0.657**	**0.857**	**0.744**

5. Conclusions and Future Work

This study provides a useful look at predictive risk modeling for future high-cost patients using a real-world data set. Results indicate that non-random sampling helps balance the challenges resulting from the skewed nature of health care cost data. Since there are multiple studies in this domain with a variety of predictors, techniques and goals, it is difficult to compare results. However, studies with similar performance metrics indicate that these results show an improvement, primarily due to the use of non-random sampling [12, 14]. Further, we compared many classification algorithms for this task and found that a select few work equally well with our data. Though we find that it is hard to choose between these algorithms, results indicate to future users a handful of appropriate classification techniques for this task. Our proposed approach for predictive modeling creates a model by

learning from the data and can therefore be adapted suitably to varied data sets. In addition, the threshold for high-cost patients is tunable and can be varied depending on the goals of a study. Comparisons using different combinations of disease-related information from various utilization classes throw up some surprises. Using disease-related information from all utilization classes is undoubtedly useful but inpatient information could prove just as useful. More importantly, demographic information alone provides a viable starting point for predictive modeling when prior data is unavailable. Additionally, we find that effective predictive models can be created without patient visit data. These results provide useful pointers regarding the selection of features appropriate for risk modeling. All these taken together signify the flexibility of our approach for predictive risk modeling and the benefits that can be obtained from such analyses.

Though our approach to balance training data using non-random sampling is intuitive, it has been found that better results could be obtained when the minority class was over-represented in the training data [5]. Hence, tweaking the sampling method is one possible direction to improve performance. The parameters of the classification algorithms can also be tuned further to improve performance. In addition, the possibility of employing cost-sensitive learning algorithms and outlier detection techniques could also be evaluated.

Predictive risk modeling is a useful technique with practical application and high impact. We provide a promising approach that is beneficial, resilient 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. Also, 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. Working with key partners and data owners, we endeavor to provide a reasonable and patient-specific answer to this question that will have a significant impact on cost containment in the health care industry.

6. References

[1] T. Bodenheimer, "High and Rising Health Care Costs. Part 1: Seeking an Explanation", *Ann Intern Med*, Nov 2005, pp. 847-854.

[2] M.L. Berk and A.C. Monheit, "The Concentration of Health Care Expenditures, Revisited", *Health Affairs*, Mar/Apr 2001, pp. 9 -18.

[3] D. Zhang and L. Zhou, "Discovering Golden Nuggets: Data Mining in Financial Application", *IEEE Trans. Sys. Man Cybernet. C Appl. Rev.*, Nov 2004, pp 513-522.

[4] N.V. Chawla, N. Japkowicz, and A. Kolcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets", *ACM SIGKDD Explorations Newsletter*, Jun 2004, pp 1-6.

[5] G.M. Weiss and F. Provost, *The Effect of Class Distribution on Classifier Learning: An Empirical Study*, tech report ML-TR-44, Dept. Computer Science, Rutgers University, Aug 2001.

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

[7] M. Maloof, "Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown", *ICML Workshop Learning From Imbalanced Datasets II*, 2003.

[8] G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data", *ACM SIGKDD Explorations Newsletter*, Jun 2004, pp 20-29.

[9] C. Drummond and R.C. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling", *ICML Workshop Learning From Imbalanced Datasets II*, 2003.

[10] A. Estabrooks, T. Jo, and N. Japkowicz, "A Multiple Resampling Method For Learning From Imbalanced Data Sets", *Computational Intelligence*, Feb 2004, pp 18-36.

[11] P. Diehr, D. Yanez, A. Ash, M. Hornbrook, and D.Y. Lin, "Methods For Analysing Health Care Utilization and Costs", *Ann Rev Public Health*, May 1999, pp 125-144.

[12] R.T. Meenan, M.J. Goodman, P.A. Fishman, M.C. Hornbrook, M.C. O'Keeffe-Rosetti, and D.J. Bachman, "Using Risk-Adjustment Models to Identify High-Cost Risks", *Med Care*, Nov 2003, pp 1301-1312.

[13] J.A. Fleishman, J.W. Cohen, W.G. Manning, and M. Kosinski, "Using the SF-12 Health Status Measure to Improve Predictions of Medical Expenditures", *Med Care*, May 2006, pp I-54-I-66.

[14] R.T. Anderson, R. Balkrishnan, and F. Camacho, "Risk Classification of Medicare HMO Enrollee Cost Levels using a Decision-Tree Approach", *Am J Managed Care*, Feb 2004, pp 89-98.

[15] Y. Zhao, A.S. Ash, R.P. Ellis, J.Z. Ayanian, G.C. Pope, B. Bowen, and L. Weyuker, "Predicting Pharmacy Costs and Other Medical Costs Using Diagnoses and Drug Claims", *Med Care*, Jan 2005, pp 34-43.

[16] A.J. Perkins, K. Kroenke, J. Unutzer, W. Katon, J.W. Williams Jr., C. Hope, and C.M. Callahan, "Common comorbidity scales were similar in their ability to predict health care costs and mortality", *J Clin Epidemiology*, Oct 2004, pp 1040-1048.

[17] J.F. Farley, C.R. Harrdley, and J.W. Devine, "A Comparison of Comorbidity Measurements to Predict Healthcare Expenditures", *Am J Manag Care*, Feb 2006, pp 110-117.

[18] I.H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques, 2nd Edition*, Morgan Kaufmann, San Francisco, 2005.