Health Care Manage Sci (2008) 11:353–358
DOI 10.1007/s10729-008-9054-y

Detecting hospital fraud and claim abuse through diabetic outpatient services

Fen-May Liou · Ying-Chan Tang · Jean-Yi Chen

Received: 3 July 2007 / Accepted: 8 January 2008 / Published online: 19 January 2008
© Springer Science + Business Media, LLC 2008

Abstract Hospitals and health care providers tend to get involved in exaggerated and fraudulent medical claims initiated by national insurance schemes. The present study applies data mining techniques to detect fraudulent or abusive reporting by healthcare providers using their invoices for diabetic outpatient services. This research is pursued in the context of Taiwan's National Health Insurance system. We compare the identification accuracy of three algorithms: logistic regression, neural network, and classification trees. While all three are quite accurate, the classification tree model performs the best with an overall correct identification rate of 99%. It is followed by the neural network (96%) and the logistic regression model (92%).

Keywords Medical insurance fraud · National health insurance · Diabetes mellitus · Data mining · Logistic regression · Neural networks · Classification trees

F.-M. Liou (✉) · J.-Y. Chen
Graduate Institute of Business Management,
Yuanpei University Hsinchu,
306, Yuanpei St.,
Hsin Chu 300, Taiwan
e-mail: mayliou@mail.ypu.edu.tw

J.-Y. Chen
e-mail: pon11290831@yahoo.com.tw

Y.-C. Tang
Graduate Institute of Business Management,
National Chiao Tung University Taipei,
118, Zhongxiao W. Rd.,
Taipei 100, Taiwan
e-mail: etang@mail.nctu.edu.tw

1 Introduction

Healthcare fraud and abuse are of major concern in many countries, in some cases costing public and private financial institutions billions of dollars (e.g., [1–3]). A growing number of healthcare insurers are using data mining tools to spot and track offenders [4]. Over the past decade or so in Taiwan, ever since its government mandated a national health care scheme, medical expenditures and utilization rates have soared. The new regime also instigated a drastic proliferation in reports of chronic diseases.

Among chronic maladies, diabetes mellitus (DM) dominates the burden on national medical expenditures. In the United States alone, treating Type 2 diabetes costs over 100 billion dollars annually. Among elderly Americans, the disease accounts for 28% of national healthcare expenses [5]. Worldwide, medical expenditures have increased from 170 million to 4,440 million US dollars from 1967 to 1999, a growth factor of about 26 [6].

In Taiwan, diabetes has become the leading chronic disease among the elderly due to changes in dietary habits and lifestyle [7]. The disease has ranked fourth among the leading causes of death in Taiwan ever since 1987. According to the Taiwan Department of Health the number of diabetes patients enrolled in the national healthcare system had climbed to 360 thousand, 1.8% of all insured. From 2001 to 2003, however, DM-related medical claims increased from 23 to 28 billion NT$ (1 US dollar = 32 New Taiwan dollars). This amount represents 8.1% of the nation's healthcare expenses, a cost far out of line with size of the affected population. The total annual claim per diabetic patient averaged 10 million NT$ (around 0.3 million US dollars) [7, 8]. The detection of fraud and abuse thus remains an important task in cost savings. The Taiwan Bureau of National Health Insurance (BNHI) is currently paying close attention to high insurance claims, in order to determine how the total expense can be reasonably reduced.

Data mining techniques have been applied successfully to the problem of healthcare fraud detection [9, 10]. For example, the Utah Bureau of Medicaid Fraud mined millions of prescriptions, surgical operations, and courses of treatment to identify unusual patterns and uncover fraud [11]. The Australian Health Insurance Commission has saved tens of millions of dollars in fraudulent claims using data mining techniques. Another success story is the Texas Medicaid Fraud and Abuse Detection System, which in less than a year has recovered $2.2 million [12].

Academic researchers have proposed several techniques for detecting individual fraudulent claims. The present study demonstrates that data mining techniques are also useful for detecting fraudulent claims at the hospital level. We compare the following three algorithms: (1) logistic regression, which has been ranked second in terms of prediction accuracy among 32 classification algorithms [13]; (2) neural networks, which are popular in many research areas; and (3) classification trees (C5.0), an ID3 (Interactive Dichotomizer 3) classification algorithm developed by Quinlan [15].

In Taiwan's national health insurance (NHI) system, payments to providers are divided into two portions: a "co-payment" provided by the beneficiary, and the "claim" subsidized by NHI. The present study uses claims data provided by healthcare providers to train and test the detection models. Section 2 briefly reviews similar attempts in the literature, and Section 3 describes the current state of Taiwan's NHI system. Section 4 presents the characteristics of the database. Section 5 defines the three algorithms in detail, and compares their performance. Section 6 concludes.

2 Data mining applications related to fraud in public healthcare

Fraud is a comparatively rare event. Nevertheless, even a small percentage can translate into an enormous number of suspect transactions at the national level. In contrast with manual monitoring, an expensive and often ineffective solution, data mining can reduce administration costs by focusing on those cases most likely to be fraudulent [12]. Perhaps more importantly, it can discover previously unknown patterns and trends in the data [16]. Healthcare providers already use data mining techniques to analyze the effectiveness of various treatments [11–14].

Koh and Tan [12] used this approach to identify the risk factors associated with the onset of diabetes. Forgionne, Gangopadhyay and Adya [7] detected healthcare fraud in a transaction database containing provider information (name, ID, and demographic data), claim information (patient ID, procedure code, charge, billing dates, and other financial data), and payment information (deductibles, co-payments, amount covered by insurance, and actual payment dates). Yang and Hwang [17] used clinical pathways to detect previously unknown abusive claims in the BNHI dataset. Chan and Lan [18] applied fuzzy set theory and a Bayesian classification algorithm to develop an abuse detection system using the same data.

3 National Health Insurance in Taiwan

Since March 1995, all Taiwanese citizens have been required by law to use the National Health Insurance (NHI). Today, the NHI coverage rate has reached 99% [7]. NHI pays for a large portion of each beneficiary's claimed expenditures on medical care and treatment managed by contracted institutions. As medication costs are increasing rapidly, the NHI has suffered from a financial deficit for the past three years (2005–2007 [7]). Among other concerns, fraudulent claims have become of prime importance. As there are not enough trained personnel to conduct a comprehensive audit, accurate identification of possible offenders is essential.

BNHI contracts with many medical care institutions overseen by the Taiwan Department of Heath: 540 hospitals (23 academic medical centers, 80 regional hospitals, and 437 local community hospitals), 16,719 medical clinics, 3,559 pharmacies, 251 medical laboratories, and 409 other medical institutions [7, 8]. (Hereafter all will be referred to as providers.) If a provider is involved in a fraudulent claim, they are fined and their contract can even be terminated.

The growth rate of NHI medical claims in Taiwan has been phenomenal. In 2003 the total amount NT$350 billion was claimed, double the 1995 level in real value. A similar increase (106%) in claims on ambulatory care arose due to both quantity and the average cost per case. Although only 16% of these services were provided by academic medical centers and metropolitan hospitals, they account for 38% of the cost. The average cost of such claims is three to four times that reported by clinical physicians, and twice that reported by community hospitals. Academic medical centers have the highest average cost per case, and the highest average cost per diem on inpatient services [7, 8].

In 2002, 44.4% of medical claims related to diabetes mellitus (DM) were for ambulatory care; the other 55.6% were for inpatient care. The average claim for a diabetic patient was about 4.3 times that for a non-diabetic patient [8]. If DM is not properly treated and controlled, it can lead to other severe conditions such as blindness, renal failure, diabetic retinopathy, nerve damage, cardiovascular problems, stroke, and peripheral vascular diseases. Oral drugs for diabetes patients promote insulin secretion (Sulfonylureas, Nateglinide and Repaglinide), suppress hepatic glucose production (Biguanides), delay intestinal digestion and absorption of carbohydrates (α-glucosidase inhibitors) [19]. DM patients also need medication to control cholesterol and blood pressure. Among adults diagnosed with diabetes, 53% only take oral medication, 19% take only insulin, 12% take both insulin and oral medication, and 15% do not take either. Self-management education is therefore integral to their medical care [20, 21].

According to BNHI reports, fraudulent and abusive claims have increased over time. The number of penalized hospitals and clinics rose from 790 in 2002 to 1,634 in 2004. In 2004, 854 institutions were fined for attempting to double their medical payments: 484 were charged to repay the difference, 162 had their contracts suspended (for one to three months), and 11 had their contracts terminated. Eradicating this waste and abuse led to savings of 140 million New Taiwan dollars [7, 8].

4 Sample data

This study defines "diabetic patients" as those for whom diabetes has been diagnosed as a principal or secondary disease (coded as A181 or A250.XX). A random sample was created by the National Health Research Institute using the NHI Database, which contains 1,050,979 diabetic patients and 17,668 healthcare providers. The sample of providers involved with fraudulent claims is simply defined as those whose contracts were terminated. Four hospitals and clinics in the sample were punished in this way, three of which (associated with a total of 189 fraudulent claims) are located in areas governed by the BNHI Central Branch. Our study only uses Central Branch healthcare providers to build the fraud detection model. Among other providers, this category contains 1,275 contracted hospitals in good standing. The database generated includes information on the diabetes patients, their diagnoses, the claims, and the healthcare providers that submitted the claims.

According to the Health Insurance Association of America [22] most medical frauds are associated with the diagnosis (43%) and billing services (34%). For the present study, we select nine expense-related variables for use in the detection models. All have previously been found useful in detecting fraudulent cases [17]. To account for differences in scale, each variable is averaged over all the cases handled by a given provider. The mean and standard deviation (over all providers) for each variable are presented in Table 1, separately for the normal providers and the three providers involved in fraudulent activities. All the three data mining procedures follow the default training-test procedure given by Clementine. The test set results are reported.

5 Data mining techniques

The SPSS application Clementine 7 is used to implement three data mining algorithms: logistic regression, a neural network, and a classification tree. All the algorithms perform better if they are trained on an evenly balanced (i.e., between fraudulent and non-fraudulent cases) dataset. This can be achieved either by duplicating the fraudulent data or reducing the non-fraudulent sample until the two populations have achieved the requested ratio [23]. Clementine 7 provides an option in the Distribution Node to generate frequency histograms. The frequency of selection from the fraudulent and non-fraudulent samples was weighted to ensure a balanced distribution between these two groups. The remainder of this section describes the theory and performance of the individual detection models.

5.1 Logistic regression

Logistic regression is a nonlinear method for modeling binary dependent variables, one that has proven very robust in a number of medical domains [24–29]. The classification variable can only have two values, which might be defined as true/false (for a model) or success/failure (for a treatment). The correlation between probability (P) and a vector of influence factors (X) can be stated as

$$P = \frac{e^{f(X)}}{1 + e^{f(X)}}. \tag{1}$$

P stands for the probability that a given institution is operating lawfully. If the logistic relationship between P and X is valid, then the probability that an institutions is operating unlawfully can written as:

$$1 - P = \frac{1}{1 + e^{f(X)}} \tag{2}$$

Combining Eqs. 1 and 3 yields the "odds ratio" for fraudulent and non-fraudulent institutions:

$$odds = \frac{p}{1-p} = e^{f(X)} \tag{3}$$

The natural logarithm of the odds ratio is then a linear function of the influence factors (X):

$$\text{logit}(P) = \log_e\left[\frac{P}{1-P}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k,$$

The logistic regression function has the advantage of being easily interpreted. (As each coefficient β_k shows the effect of a one-unit change in its corresponding variable on the logarithm of the predicted odds ratio, the variables with a larger coefficient are more useful in detecting fraudulent cases.) A maximum likelihood method is usually used to find the model that best distinguishes the two groups. A

Table 1 Descriptive statistics for normal and fraudulent hospitals

Variable (per case)	Normal Hospital		Fraudulent Hospital	
	Mean	SD	Mean	SD
Average days of drug dispense	7.72	5.60	7.39	1.50
Average drug cost	221.63	274.06	208.25	88.13
Average consultation and treatment fees	358.71	176.88	259.58	113.69
Average diagnosis fees	265.42	42.93	265.00	43.46
Average dispensing service fees	24.48	8.13	30.01	11.21
Average medical expenditure	548.04	408.33	584.75	145.92
Average amount claimed	487.99	394.69	511.81	131.97
Average drug cost per day	28.81	27.79	33.82	10.07
Average medical expenditure per day	134.37	92.29	173.13	73.56

claim was associated with the value “0” if regular, and “1” if irregular.

Stepwise logistic regression was performed on each variable individually to identify the most effective factors (Table 2). Eight of the nine variables were found to have significant predictive power (“average medical expenditure” does not). Those eight detectors were subsequently used to create a full logistic regression model. The detection rate on fraudulent hospitals is 100% (three out of three are detected), while the correct identification rate for normal hospitals is 84.6% (Table 3). The correct identification rate for the whole sample is 92.2%.

5.2 Neural networks

A neural network emulates the human brain to classify and predict data. It consists of a simple network of several artificial “neurons”, or nodes, some of which receive scalar data from other nodes and transform the information to a single output signal. The interconnections are weighted, and these weights are modified as the network operates on training data. A typical neural network for data classification consists of three or more layers: the input layer, a hidden layer, and an output layer. Nodes in the hidden layer receive a weighted sum of the input variables, and transform that sum to an output signal using some kind of threshold function (typically a step function or sigmoid) [30, 31]. The output layer (often a single node) receives a weighted sum of the hidden layer’s outputs, and converts it to a classification signal in the same way. Neural networks can establish a relationship between the input and output data without external guidance, and are effective in many cases. The self-organizing map of weights in some sense simulates the biological learning process of neural systems [32, 33].

Neural networks do not specify the significance of individual variables, however, which may be considered a disadvantage. Clementine neural nets have sensitivity analysis, which shows which variables are more important for the classification. The result indicates the ranking of each variable’s relative importance in classifying the data as: average dispensing service fees, average diagnosis fees, average medical expenditure per day, average days of drug dispense, average drug cost per patient, average drug cost per patient per day, average consultation and treatment fees, average medical expenditure, and average amount claimed. For the neural network algorithm, the correct identification rates for the whole test sample and normal hospitals in particular are 95.73 and 91.47% respectively (Table 3).

Table 2 Results of stepwise logistic regression

Variable (per case)	Log Likelihood of the Reduced Model	p Value at 1% Level of Significance
Intercept	2,562.454	0.000***
Average days of drug dispense	1,449.564	0.000***
Average drug cost	1,448.026	0.000***
Average consultation and treatment fees	1,439.871	0.012***
Average diagnosis fees	2,053.640	0.000***
Average dispensing service fees	2,041.844	0.000***
Average medical expenditure	1,765.307	0.000***
Average amount claimed	1,794.374	0.000***
Average drug cost per day	1,508.997	0.000***
Average medical expenditure per day	1,433.659	0.748

Table 3 Prediction results of three data mining algorithms

Samples	Normal Result	Fraud Result	Total	Accuracy (%)	Average Accuracy
Logistic regression					=(1,069 1,267)/2,534
Normal hospitals	1,069	198	1,267	84.60	
Fraud hospitals	0	1,267	1,267	100.0	=92.18%
Neural network					=(1,159 1,267)/2,534
Normal hospitals	1,159	108	1,267	91.47	
Fraud hospitals	0	1,267	1,267	100.0	=95.73%
Classification tree					=(1,251 1,267)/2,534
Normal hospitals	1,251	16	1,267	98.73	
Fraud hospitals	0	1,267	1,267	100.0	=99.37%

5.3 Classification tree

Classification tree algorithms are used to predict the membership of cases defined by a categorical dependent variable. Each "branch node" of the tree (the first one is the "root"), partitions the data into two or more sub-branches. The classification procedure stops when the bottom level of "leaf nodes" defining the categories is reached [14]. After building a classification trees based on the training set, the rules or patterns in the data are obvious and can be implemented in a straightforward detection algorithm. Each possible path from the root node to the leaf node represents a sequence of classification rules. Two of the rules encountered on the way to the "fraud" leaf node are stated explicitly below, by way of illustration.

(1) Rule 1:

IF "average days of drug dispense" <=6.0070958, THEN "contract status" = regular → stop at this branch
IF "average days of drug dispense" >6.0070958, THEN "contract status" = irregular → move on to the next node

(2) Rule 2:

IF "average medical expenditure" <=416.23779, THEN "contract status" = regular → stop at this branch
IF "average medical expenditure" >416.23779, THEN "contract status" = irregular → move on to the next node
:
:

The sequence continues until an optimal prediction is obtained.

For the classification tree algorithm, the detection rate on fraudulent hospitals is 100%. The correct identification rates for the whole dataset and normal hospitals are 99.30 and 98.73% respectively (Table 3).

5.4 Comparison the results of the three forecasting models

All three algorithms achieved a correct identification rate of 100% for fraudulent institutions. However, their accuracies on the normal providers are different. The classification tree model has the smallest error rate (1%) in classifying normal providers; it is followed by the neural network model (9%) and the logistic regression model (15%).

6 Conclusions

This study employed three data mining techniques—logistic regression, neural networks, and classification trees—to detect fraudulent healthcare providers in the Taiwan NHI database based on their submitted claims. All three approaches detect the fraudulent and abusive medical care institutions. In terms of overall accuracy, the classification tree is superior to the logistic regression and neural network models. The high rates of correct identification indicate that the selected variables can identify hospitals submitting irregular medical claims.

Assuming a one hundred percent identification rate for irregular institutions in practice, the algorithm that features the lowest ratio of wrongly identified normal institutions could be implemented at the lowest cost. In light of this fact, the C5.0 tree is the optimal detection model. It specifies the following sequence of variables, from most to least important: average length of drug treatment (days), average total medical expenditure, and average consultation and treatment fees.

The main limitation of this study is that the sample of fraudulent providers includes only three hospitals whose contracts were terminated by BNHI. Fraudulent and abusive institutions that padded their bills but were not penalized so severely were not included in the sample due to data limitations.

In Taiwan, healthcare providers submit more than 30 millions claims per month. The cost of reviewing every case would be exorbitant without a modern screening system. Rather than screening for fraud at the level of individual claims, we propose that providers can be classified using data on their diabetic outpatient services. Since these claims account for a significant fraction of their total expenses, the statistical properties of related financial variables can be used to identify institutions which are more likely to be participating in fraud.

Acknowledgements The authors thank the National Science Council of the Republic of China for financially supporting this research (NSC 95–2416-H-264 -010).

References

1. Lassey M, Lassey W, Jinks M (1997) Health care systems around the world: characteristics, issues, reforms. Prentice-Hall, NJ
2. Hong S, Weiss S (2001) Advances in predictive models for data mining. Pattern Recog Lett 22(1):55–61
3. Bolton R, Hand D (2002) Statistical fraud detection: a review. Statist Sci 17(3):235–249
4. Koh H, Tan G (2005) Data mining applications in healthcare. J Healthc Inf Manag 19(2):64–72
5. Ratner R (1998) Type 2 Diabetes Mellitus: the grand overview. Diabetic Med 15(S4):S4–S7
6. Guisseppi F, Gangopadhyay A, Adya M (2000) Intelligent data mining system to detect healthcare fraud. In: Armoni A (ed) Healthcare information systems: challenges of the new millennium. Hershey, Idea Group Publishing, PA
7. National Health Insurance Bureau (2000–2004) National Health Insurance Statistics
8. Bureau of National Health Insurance (2004) Report on quality of medicare for diabetes mellitus under National Health Insurance
9. Sparrow M (1996) License to steal. Westview Press, Boulder, CO
10. Long J, Irani E, Slagle J (1991) Automating the discovery of causal relationships in a medical records database. In: Piatestsky-Shapiro G, Frawley W (eds) Knowledge discovery in database. AAAI Press, Menlo Park, CA
11. Milley A (2000) Healthcare and data mining. Health Manag Technol 21(8):44–47
12. Koh H, Gerald T (2005) Data mining applications in healthcare. J Healthc Inf Manag 19(2):54–72
13. Lim T, Loh W, Shih Y (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Mach Learn 40(3):203–229
14. Chae Y, Seung H, Kyoung W, Dong H (2001) Data mining approach to policy analysis in a health insurance domain. Med Inf 62(2):103–111
15. Quinlan J (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Mateo, CA
16. Kincade K (1998) Data mining: digging for healthcare gold. Ins Technol 23:IM2–IM7
17. Yang W, Hwang S (2006) A process- mining framework for the detection of healthcare fraud and abuse. Expert Syst Appl 31(1): 56–68
18. Chan C, Lan C (2001) A data mining technique combining fuzzy sets theory and bayesian classifier-an application of auditing the health insurance fee. Proceedings of the International Conference on Artificial Intelligence, IC-AI'2001, Las Vegas, USA: 402–408
19. Bloomgarden Z (2002) The epidemiology of complications. Diabetes Care 25(5):924–932
20. American Diabetes Association (2003) Report of the expert committee on the diagnosis and classification of diabetes mellitus. Diabetes Care 26(Suppl 1):S5–S20
21. American Diabetes Association (2004) Standards of medical care in diabetes. Diabetes Care 27(1):S15–S35
22. Health Insurance Association of America (1993) Update Source Book of Health Insurance Data
23. Nisbet R (2006) Data mining tools: which one is best for CRM? Part 3, BI Report, March (2006), available at http://www.dmreview.com/editorial/dmreview/print_action.cfm?articleId=1049954. Accessed on March 9, 2007
24. Young M, Seung H, Dyoung W, Dong H, Sun H (2001) Data mining approach to policy analysis in a health insurance domain. Int J Med Inf 62:103–111
25. Dasgupta C, Despensa G, Ghose S (1994) Comparing the predictive performance of a neural network model with some traditional market response models. Int J Forecast 10(2):235–244
26. Fish K, Barnes J, Aiken M (1995) Artificial neural networks: a new methodology for industrial market segmentation. Industrial Market Manag 24(5):431–438
27. Hruschka H (1993) Determining market response functions by neural network modeling: a comparison to econometric techniques. Eur J Oper Res 66(1):27–35
28. Cabena P, Hadjinian P, Stadler J, Zanasi A (1998) Discovering data mining from concept to implementation. Prentice Hall PTR, Upper Saddle River, NJ
29. McKee T, Lensberg T (2002) Genetic programming and rough sets: a hybrid approach to bankruptcy classification. Eur J Oper Res 138:436–451
30. Widrow B, Rumelhart D, Lehr M (1994) Neural networks: applications in industry, business and science. Commun ACM 37(3):93–105
31. Wilson R, Sharda R (1994) Bankruptcy prediction using neural networks. Decis Support Syst 11(5):545–557
32. Wu J (1994) Neural networks and simulation methods. Marcel Dekker Inc., NY
33. Haykin S (1994) Neural network: a comprehensive foundation. Prentice Hall PTR, Upper Saddle River, NJ

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.