Expert Systems with Applications 39 (2012) 7441–7450

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A scoring model to detect abusive billing patterns in health insurance claims

Hyunjung Shin [a], Hayoung Park [b,*], Junwoo Lee [a], Won Chul Jhee [c]

[a] Department of Industrial and Information Systems Engineering, Ajou University, San 5 Woncheon-dong, Yeongtong-gu, Suwon 443-749, Republic of Korea
[b] Technology Management, Economics, and Policy Graduate Program, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-744, Republic of Korea
[c] Department of Industrial and Information Engineering, Hongik University, 72-1 Sangsu-dong, Mapo-gu Seoul 121-791, Republic of Korea

ARTICLE INFO

Keywords:
Health insurance claims
Medical abuse detection
Fraud detection
Degrees of anomaly
Data mining

ABSTRACT

We propose a scoring model that detects outpatient clinics with abusive utilization patterns based on profiling information extracted from electronic insurance claims. The model consists of (1) scoring to quantify the degree of abusiveness and (2) segmentation to categorize the problematic providers with similar utilization patterns. We performed the modeling for 3705 Korean internal medicine clinics. We applied data from practitioner claims submitted to the National Health Insurance Corporation for outpatient care during the 3rd quarter of 2007 and used 4th quarter data to validate the model. We considered the Health Insurance Review and Assessment Services decisions on interventions to be accurate for model validation. We compared the conditional probability distributions of the composite degree of anomaly (CDA) score formulated for intervention and non-intervention groups. To assess the validity of the model, we examined confusion matrices by intervention history and group as defined by the CDA score. The CDA aggregated 38 indicators of abusiveness for individual clinics, which were grouped based on the CDAs, and we used the decision tree to further segment them into homogeneous clusters based on their utilization patterns. The validation indicated that the proposed model was largely consistent with the manual detection techniques currently used to identify potential abusers. The proposed model, which can be used to automate abuse detection, is flexible and easy to update. It may present an opportunity to fight escalating healthcare costs in the era of increasing availability of electronic healthcare information.

1. Introduction

Increasing healthcare costs have burdened the economies of almost every developed and developing country, and the problem is worsening with an aging population and advancing health technology (Organisation for Economic Co-operation, 2008; 2009). Ongoing efforts against medical abuse and fraud include steps to reduce inappropriate use of healthcare funded by third-party payers, but the process is costly (Center for Medicare, 2008a; Feldman, 2001; Pontell, Jesilow, & Geis, 1982; Rai, 2001; Shane, 2000). Various estimates suggest that the magnitude of the problem, measured as a percentage of the healthcare budget, would range from 3 to 10% in the United States; however, key statistics for South Korea are unavailable. The Improper Medicare FFS Report indicated that 3.7% of US Medicare payments were inappropriate, which amounted to 10.2 billion USD in FY 2007, and in its 2007 crimes report, the US Federal Bureau of Investigation (FBI) estimated the figure to be as high as 10% (Center for Medicare, 2008b; National Health Care Anti-Fraud Association, 2009). The FBI identified healthcare fraud schemes such as billing for unprovided services, upcoding services and items for higher payments, submitting duplicate claims, unbundling services that should be billed as a single item, providing medically excessive and unnecessary services, and kickbacks (US Federal Bureau of Investigation, 2009). Medical abuse and fraud compromise both healthcare costs and quality. They also harm honest and ethical healthcare providers.

The detection of abusive and fraudulent practice in healthcare is difficult because uncertainties inherent in medical practices result in variable care processes (Eisenberg, 2002; Henderson, 2009). Therefore, medical experts must review each case, which can be time consuming and expensive. Advances in information technology and digitization of healthcare information, such as electronic medical records, bills, and claims, opened a new venue for efficient and effective medical abuse and fraud detection. Data mining and machine learning technologies have been widely used for fraud detection and auditing in the auto and life insurance, banking, credit card, and mortgage industries, and since the late 1990s, similar efforts have been made in healthcare (Hager et al., 2006; Li, Huang, Jin, & Shi, 2008).

Interest in fraud detection research has been gaining strength mainly in developed countries, and the scope of the research is expanding (Phua, Lee, Smith, & Gayler, 2005). Fraud detection models are most often cited for the national security, industrial

* Corresponding author. Tel.: +82 2 880 2575; fax: +82 2 886 8220.
E-mail addresses: shin@ajou.ac.kr (H. Shin), hayoungpark@snu.ac.kr (H. Park), neo100in@ajou.ac.kr (J. Lee), jhee@hongik.ac.kr (W.C. Jhee).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.105

information security, credit card, e-commerce, insurance, and telecommunication industries. Traditional statistical methods, data mining algorithms, and new machine-learning methods are used in the detection models. Although various algorithms are applied depending on the nature of the problem aimed at in the healthcare domain, neural network algorithms which result in superior performance and decision tree algorithms which render easy to understand results are two of the most popular methodologies (Bonchi, Giannotti, Mainetto, & Pedreschi, 1999; He, Hawkins, Graco, & Yao, 2000; Major & Riedinger, 2002).

Newer technologies are introduced to detect fraud and recently researchers have combined multiple methodologies. For example, some researchers used fuzzy logic in medical claims assessment, a combination of heuristic searches and the activity rule group for fast data review, and neural network algorithms to automatically classify claims information (Brockett, Xia, & Derrig, 1998; Cox, 1995). Also a combination of a neural network and a genetic algorithm along with application of Naive Bayes was attempted in the assessment of fraud in claims (Viaene, Richard, & Dedene, 2005). Koh and Tan (2005) introduced cases of data mining application, including for fraud and abuse detection, in the broad spectrum of healthcare management.

Attempts to apply data mining methodologies at the national and state level can be found in the previous studies as well. The Health Insurance Commission of Australia, which administrates the Medicare program for the Australian federal government, used online unsupervised learning algorithm based on finite mixture model to detect outliers in the utilization of pathology services (Yamanishi, Takeuchi, Williams, & Milne, 2004) and a combination of two neural network algorithms, the multi-layered perceptron (MLP) and Self Organizing Map, to identify abnormal patterns from the practice profiles of general practitioners (He, Wang, Grac, & Hawkins, 1997). The National Health Insurance (NHI) program of Taiwan developed disease-specific clinical pathways to identify fraudulent claims. The detection model, based on a process mining framework, automatically and systematically identified practices that deviated from the pathways, which could indicate abuse and fraud (Yang & Hwang, 2006). Also the NHI attempted to apply a model that combined fuzzy sets theory and a Bayesian classifier to a claims audit (Chan & Lan, 2001). The application of MLP neural networks in medical abuse and fraud detection enabled a Chilean private health-insurance company to install a real time-based detection process that brought considerable savings to the company (Ortega, Figueroa, & Ruz, 2006).

Since 1989, the entire population of South Korea (49.5 million) has been covered by a uniform insurance policy administered by the National Health Insurance Corporation (NHIC), except for approximately 3.6%, who are covered through the medical aid program funded by the general tax. Physicians and hospitals are reimbursed based on a fee-for-service mechanism, based on a fee schedule predetermined annually by the government. Although fees are strictly regulated by the government, the system is vulnerable to providers' abusive utilization and billing behavior, which causes unnecessary increases in healthcare costs. The insurer instituted a prepayment claims review and audit process to prevent improper utilization from reimbursement, and the Health Insurance Review and Assessment Services (HIRA) is dedicated to claims review and audit. Despite all the concerted efforts by the NHIC and HIRA, between 1990 and 2007 the health insurance budget expanded at an average annual rate of 16% (Organisation for Economic Co-operation & Development, 2009; Health Insurance Review & National Health Insurance Corporation, 2008).

The magnitude of information reviewed and audited by the 1700 employees of the HIRA is enormous and growing fast. In 2007, a total of 968 million claims were submitted to the HIRA for reimbursement, and nearly 37% of them were outpatient claims submitted by clinics (or physician practices). The average annual growth rate of claims between 2000 and 2007 was 13% and the average size of an outpatient claim was 1.9 KB (8.7 KB for inpatient claims). At first, every claim was manually reviewed to determine the amount of reimbursement, but organizational expansion proved politically infeasible, HIRA staff quickly realized that the practice was unsustainable. HIRA management saw opportunities in information technology and focused a strategy to simultaneously enhance the effectiveness and efficiency of the organization.

Claims were digitized starting in 1994 and the Electronic Data Interface (EDI) based billing system was introduced in 1996. As of 2008, 97% of 78,410 clinics, hospitals, and pharmacies submitted electronic claims. A data warehouse (DW) was built in 2003 so reviewers were better equipped with knowledge extracted from claims information. The size of the DW, which keeps 5-year claims information, was 142 TB in 2008. In another initiative, launched in 2002 to capitalize on the majority of claims being in electronic forms, reviewers focused on potential abusers to prevent waste, the Comprehensive Intervention Program, instead of post-utilization reviews. Under this program, machines do most of the post-utilization reviews on outpatient records while the reviewers undertake manual review to detect and educate and communicate with the small percentage of providers with abusive utilization behavior. Twenty-six thousand clinics submitted 67% of the outpatient claims, but the contents tended to be simple compared to the inpatient and outpatient claims submitted by hospitals.

Reviewers manually selected clinics based on approximately 180 indicators routinely computed in the DW using individual providers' claims data. Some examples of the indicators that comprise the case-mix adjusted costliness indices (CIs) for total charges and charges for categories of services such as IV, procedures, antibiotics, expensive medications, and lab work. The case-mix adjusted indicators of intensities of utilization also include data on the numbers of prescription medications and the days medications are prescribed as well. All the information characterized by about 180 indicators is difficult to amalgamate manually, therefore providers were selected for further investigations based on rankings of individual indicators regardless of the significance of the problem found. For example, a provider ranked in the top 3‰ for one indicator but below 50‰ in all other indicators could be selected, but one that ranked in the top 10‰ for all indicators may not be selected. The selection process based on these rankings shows obvious flaws. Furthermore, the manual selection of clinics with abusive billing patterns grew increasingly complicated because new treatments and medicines increased the information considered in the selection. The process has been criticized for lacking rationality, consistency, and interpretability.

We formulated a model that detects healthcare providers who show a pattern of abusive behavior in the provision of outpatient care. The proposed model was designed to automatically process large amounts of information contained in healthcare insurance claims and to generate an index that can be used to decide whether further investigation of the practitioner for subsequent intervention is warranted. We also applied the decision tree method to create clear explanations about the characteristics that make a provider a suspect of abuse.

2. Material and methods

2.1. Modeling

The proposed model is designed through two phases of modeling: scoring and segmentation. The scoring model quantifies the degree of abusiveness in a provider's billing pattern and the segmentation model groups providers based on the resulting

scores. Providers with top-tier scores are profiled further in the segmentation model to determine the reasons their scores are higher than those of other providers.

2.1.1. Scoring

Finding providers with abusive utilization patterns is challenging because no clear-cut definition about undesirable utilization has been articulated and screening out suspicious providers by checking their claims submitted to third-party payers in a fee-for-service based payment system is difficult. Claims contain much information about various types of medical services rendered, such as medications, injections, diagnostic imaging and tests, and procedures. Indicators derived from this information are likely to contain partly independent and partly complementary information about abusive behavior. We attempt to mimic the ways a human examiner processes information to identify providers with undesirable billing patterns, which are assessing the magnitude of problems and amalgamating the results of the assessment.

For a specific indicator, the examiner does an a priori search of the top-ranked providers while the low-ranked ones (often below the average score) are excluded from the examination. If the investigator only considers rank, the provider with higher amounts of medical utilization will be identified without the magnitude of problems. The *degree of anomaly* (DA) is defined as follows:

$$\mathrm{DA}(x_{ij}) = \exp\left(\left[\frac{\max(x_{ij}-\mu_j,0)}{\sigma_j}\right]^2\right) \tag{1}$$

where x_{ij} is the value of a specific indicator j of provider i, and μ_j and σ_j are the average and the standard deviation of the corresponding indicator j, respectively. The term $\max(x_{ij}-\mu_j, 0)$ implies that the value of the index for all providers with indicator scores below the average value is set to the minimum value of zero, removing them from consideration for further review. With a large number of providers to be reviewed, one can eliminate below-average scores (nearly one-half of the sample) alleviating the time and memory necessary for calculations. Furthermore, the index defined in Eq. (1) assigns higher weights to scores far above the average by taking the exponential function.

Because we are interested in understanding how a human reviewer amalgamates information drawn from DAs that indicate diverse aspects of abusive patterns, we departed from the previous approach in which the respective ranks of indicators were individually considered and defined the *composite degree of anomaly* (CDA). Several examples of utilization patterns may be studied to find an approach for amalgamation of information, and Fig. 1 shows examples of providers with different utilization patterns. For simplicity, five indicators are considered in the examples: rate of injection, CI of total charges, rate of costly prescriptions, number of medication days per claim, and number of medications per prescription. We use radial diagrams to assess each provider's utilization profile in the five types of indicators simultaneously, and each axis of the diagram represents a scaled indicator relative to its range. The vertex of an inner shaded regular pentagon stands for the average of the corresponding indicator. Fig. 1(a) depicts a typical shape for a non-abusive provider. In contrast, both (b) and (c) likely represent profiles of abusive providers. The provider in (b) draws investigative attention because the injection-rate value is extraordinary higher compared to that of other providers. The provider in (c) does not show any high-peak in a specific indicator, but utilization across all five indicators tends to be larger than average. This type of abusive pattern is seldom detected in an indicator-by-indicator search because none of the values is pronounced.

The CDA defined in Eq. (2) attempts to aggregate the net amount of overuse across all indicators; it takes the weighted average of the individual DAs, and the aggregation also involves a decision on which indicators are most important in identifying abusive providers, which translates as the weight for each indicator:

$$\mathrm{CDA}(x_i) = \frac{\sum_{j=1}^{n} w_j \exp\left(\left[\frac{\max(x_{ij}-\mu_j,0)}{\sigma_j}\right]^2\right)}{\sum_{j=1}^{n} w_j} \tag{2}$$

where w_j indicates a weight or combination coefficient.

We can apply one of several statistical techniques to find the DA weights in Eq. (2) (Guyon & Elisseeff, 2003; Liu & Motoda, 2001). The techniques mostly rely on the inter-relationship between the input variables (indicators) and the binary output variable, which indicates the presence or absence of abusive billing patterns. However, as noted by Simborg (2008), it is hard to comprehend the magnitude of health fraud. The domain experts (eg, HIRA claims reviewers) do not have confidence about the selection of abuse patterns; therefore, we do not have information about the output variable. To accommodate this case, the CDA can be designed to be less dependent on the change of weights. Note that the index DA in Eq. (1) varies with the exponential scale of indicators whereas the composite index CDA in Eq. (2) varies with a linear scale of DAs. These characteristics imply that the individual DA score is highly significant in terms of showing abnormality but the impact of weights on the integrated CDA index is relatively trivial. A simple uniform weighting may work, however, we attempted to utilize the past selection made by HIRA domain experts despite the recognition of its imperfect qualities (we expected imperfect information to provide better outputs than would no information at all). We employed six statistical techniques — correlation analysis (ρ), logistic regression (R), *t*-test (t), entropy-reduction (E), discriminant analysis (D), and Chi-square test (χ) — to compute six different weights for each indicator, and then we computed the consensus weight, w_j, for the jth indicator by summing the six weights as presented in Eq. (3):

$$w_j = w_j^{\rho} + w_j^{R} + w_j^{t} + w_j^{E} + w_j^{D} + w_j^{\chi} \tag{3}$$

The weights can be interpreted as the strength of the association between an indicator and the domain experts' selection of clinics for intervention.

2.1.2. Segmentation

With the CDA score, reviewers with domain expertise can tailor responses toward an abusive provider based on the seriousness of practitioner behavior as measured by the index. For a provider in the top CDA scores, a reviewer may make an intensive investigative and correctional intervention; for providers in the mid-high range of the CDA, he/she may prescribe a mild intervention or simple recommendation for betterment. However, the CDA score is neither suggestive nor intuitively understandable to either examiners or providers because it does not delineate the activity that draws the suspicion nor the behaviors needed to improve. Decision trees are helpful in understanding the results (Breiman, Friedman, Olshen, & Stone, 1984; Quinlan, 1993).

After sorting providers by the CDA score, we arranged providers into several groups for *grade for degree of anomaly* (GDA) classification. The number of groups is user-specified, and the cutoff scores can be either defined by the user or determined in an equi-width or equi-frequency manner. The equi-width creates cutoffs at nearly equally spaced intervals of the CDA whereas the equi-frequency creates groups with approximately equal frequencies. The GDA discretizes the CDA, which means the continuous value of the CDA with uncountable levels and unbounded ranges is converted into an ordinal value with few levels and bounded ranges. This simplification is of practical benefit to domain experts because they can, for instance, tailor the number of the intervenient actions that does not exceed the number of grades.

Because of limitations in time and manpower, the investigation and intervention focus on a few top-scorer groups. A higher score

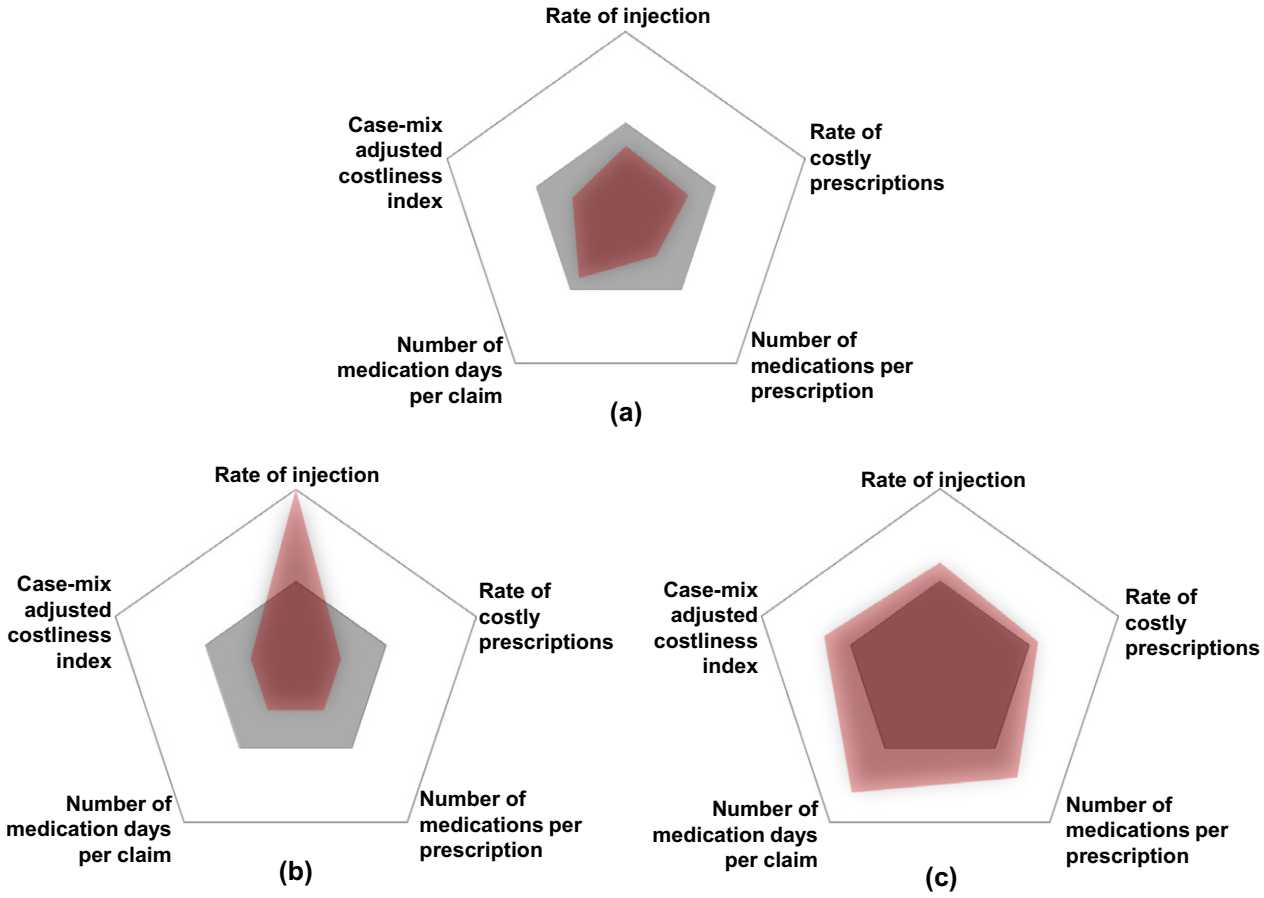

Fig. 1. Examples of the radial diagram of providers with different billing and utilization patterns.

represents a higher degree of anomaly. However, depending on the indicators that contribute the most to a high CDA score, providers in a high score group differ from each other in their abusive behavior. A set of providers with similar problematic indicators is called a *segment*. For the segmentation, we used the GDA as the target variable for the decision tree. By segregating the suspicious providers into homogeneous segments, an investigator now can tailor a proper investigative and correctional intervention for each segment.

One can gain interpretability for the CDA score by reclassifying the groups using decision trees. Fig. 2 exemplifies the case of a four-group classification: GDA(1)–GDA(4). The leaf (shaded) nodes are labeled as one of four group-grades after training. One can determine the indicator that contributed significantly for the node-split by tracing back the tree from the leaf node to the root. The providers most likely to be abusive—the ones with a very high CDA score, GDA(4)—are segregated into two segments based on the difference in their abusive utilization patterns: the CI of total charges and the rate of prescribing antibiotics. With the help of the decision tree, we see the types of anomalous utilization pattern that led a reviewer to select the provider for the intervention, and the reviewer can justify the selection and can offer instructive explanations about the decision. Furthermore, the payer can offer a relatively sophisticated decision with the automatically generated rule and can pursue an appropriate penalty or action on the specific provider.

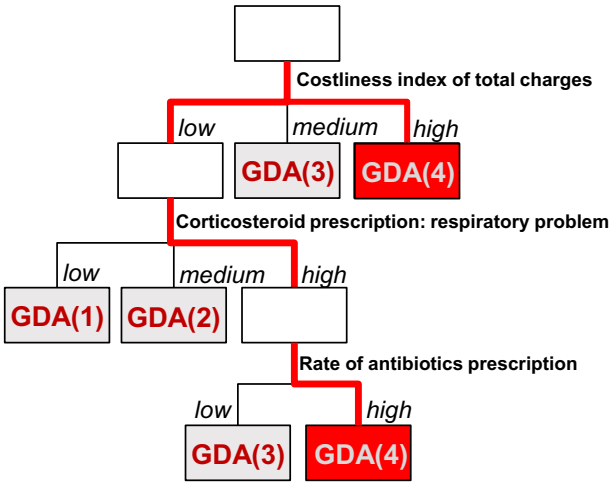

Fig. 2. Using a decision tree to gain interpretability by reclassifying the four groups GDA(1)–GDA(4).

2.2. Data and validation

Clinics, or physician practices, are required to submit a monthly claim to the HIRA for each outpatient. To receive fee-for-service based payment, the provider must include all charges, including those from multiple visits, incurred by a patient in a particular month. The claim also reveals utilization information of all different types of services and products rendered such as procedures, treatments, tests, and medications. To review claims and decide on the appropriateness of utilization and to finalize the payment, the HIRA has relied on indicators. For each clinic, the HIRA information system generates utilization and billing profiles derived from information submitted in claims as well as additional

demographic and disease characteristics of the patient group and the region. Indicators are updated every three months.

The HIRA made available the 45,000 quarterly records of 187 indicators for 28,066 clinics in general practice and 16 specialties during the last two quarters of 2007. We examined the association between the selection for intervention by the HIRA and each indicator to reduce the data to a manageable size (Matignon, 2007). We rejected the indicators in the continuous scale with r^2 improvement values of less than 0.005 and the categorical indicators with associations not significant at $\alpha = 0.2$ in the chi-square tests. The final set of the selected 38 indicators was verified by the domain experts at the HIRA. They include CIs of various charges such as those for total utilization, medications, injections, laboratory tests, and diagnostic radiology; CIs of total charges for the five most frequent diagnoses; rates of utilization of services closely monitored by the HIRA, such as antibiotics and corticosteroids; utilization of services such as visits and prescription drugs; and the detailed definition of 38 indicators is presented in the Appendix. To adjust for case-mix differences in computing CIs at the clinic level and measure the intensity of utilization, the HIRA uses the Korean Outpatient Groups, an outpatient classification system similar to the US Ambulatory Payment Classification, and it codes diagnoses by the International Classification of Disease, 10th Revision (ICD-10). About 9% (2547) of 28,066 clinics had intervention records during the study period. We performed the modeling for each of the general practice and 16 specialties that include internal medicine, surgery, obstetrics and gynecology, pediatrics, neuropsychiatry, neurology, dermatology, orthopedic surgery, neurosurgery, urology, otorhinolaryngology, ophthalmology, radiology, anesthesiology, rehabilitation medicine, and family medicine. Presented in this article is the full-scale modeling for the most common of those 17 groups: internal medicine.

We split the study data into two subsets, the 3rd and 4th quarters of 2007, and performed the proposed modeling on the 3rd quarter data and validated the model with the 4th quarter data. We used the intervention decision made by the HIRA based on the 4th quarter claims, which we considered to be accurate values in the validation. Among 3705 internal medicine clinics, 359 were selected for intervention. We compared the conditional probability distributions, p(CDA|non-intervention) and p(CDA|intervention) and examined confusion matrices by intervention decision and group as defined by the CDA score to assess the validity of the proposed CDA index. We also examined the percent of payment denied by the HIRA in the confusion matrices.

3. Results

3.1. Scoring with the composite degree of anomaly

Integrating information in 38 indicators, the CDA score quantifies the abusiveness in utilization and billing of 3705 internal medicine providers. Fig. 3 illustrates the relative magnitude of 38 weights, or combination coefficients, used to compute the CDA scores defined in Eq. (2). The upper panel presents the weights for clinics in internal medicine, and the graphs for the general practice and remaining 15 specialties are presented in the lower panel of the figure. In the case of internal medicine, the most pronounced indicators (i.e., with a weight value greater than 3) are the CI of total charges, the CIs of total charges for patients with the most and the second-most frequent diagnosis in internal medicine, the CI of charges for oral medication, the number of medications per prescription, and the rate of prescriptions with more than six medications (5.129, 3.765, 3.595, 3.502, 3.413, and 3.212, respectively). The least significant indicators, with values less than 1, include the prescription rate of costly medications, the number of visits per claim, and the CIs of charges for CT, MRI, registration, and CMI. This result indicates that, when manually selecting internal medicine clinics for thorough investigational and correctional intervention, HIRA reviewers had been more focused on (or more frequently looked into) high weight indicators. In Fig. 3, the set of indicators that reviewers focused on vary by specialty. For example, the values of the weights of the CIs of charges for CT and MRI are above the average in surgery but not in internal medicine.

3.2. Grouping into grade for degree of anomalies

Internal medicine clinics are grouped into five GDAs as an ad hoc set-up based on their CDA scores. The cutoff values of the CDA scores were roughly set according to the CDA order of the magnitude. Table 1 shows the boundaries and the distribution of clinics among the GDAs. The GDA(4) is the most suspicious or abusive group of providers, whereas the GDA(0) is the group with the least problematic utilization pattern. More than one-half of the 3705 clinics (63%) were grouped into the GDA(0) and would be the first to be excluded from an in-depth intervention if the HIRA needed to concentrate resources on the group with serious utilization problems. The claims in the 236 GDA(4) clinics (6%) revealed the most abusive utilization patterns, and the payer can expect the largest gain by correcting their utilization behavior.

Fig. 4 shows the radial diagrams of the DA scores of 38 indicators of typical providers belonging to each of the five GDA groups. The pattern of Provider A belonging to GDA(4), the diagram in the upper panel, draws investigative attention because of larger-than-average values of a number of indictors, particularly of the VI and the CI of consultation fees. However, the lower right insert presents the DA pattern of Provider E assigned to the group GDA(0). Very different from the diagram in the upper panel, it shows that the values of 38 indicators are all lower than or near averages.

3.3. Segmentation using decision trees

Because the high-scoring GDA providers show diverse patterns in abusive behavior and we want to trace the characteristics that led reviewers to choose practitioners for further intervention, we used decision trees to examine the indicators of 3705 internal medicine providers. These indicators were the input variables of the tree, and the five GDA grades were the target values in the analysis. The tree splits the root node of the 3705 providers into several children nodes by denoting the most significant indicators in the GDA classification. The input indicator in a higher level node of the tree is more important than one in a lower level. The CI of total charges was the most important indicator at the first level of the tree. The indicators at the second level were the VI, rate of injection, and number of medications per prescription. As the tree grows, the purity at the leaf nodes, measured by the proportion of providers assigned to the dominant GDA, increases and a leaf node at the lowest level in the resulting tree is a segment of the providers who are similar in their abusive behavior.

Fig. 5 shows the two segments of providers assigned to the GDA(4). Although they were all assigned to GDA(4) because of a high CDA score, abusive patterns that led the GDA assignment were different and so the internal medicine providers depicted were further assigned to two different segments. The providers A–D had high VI values even though their CI of total charges were low, whereas the providers E–K had high CI values, a high number of medications per prescription, and a high utilization of corticosteroids for joint problems indicated by principal or secondary diagnosis coded by the ICD-10 codes, M13∼M17 or M19. The segmentation result can assist claims reviewers in two ways: It

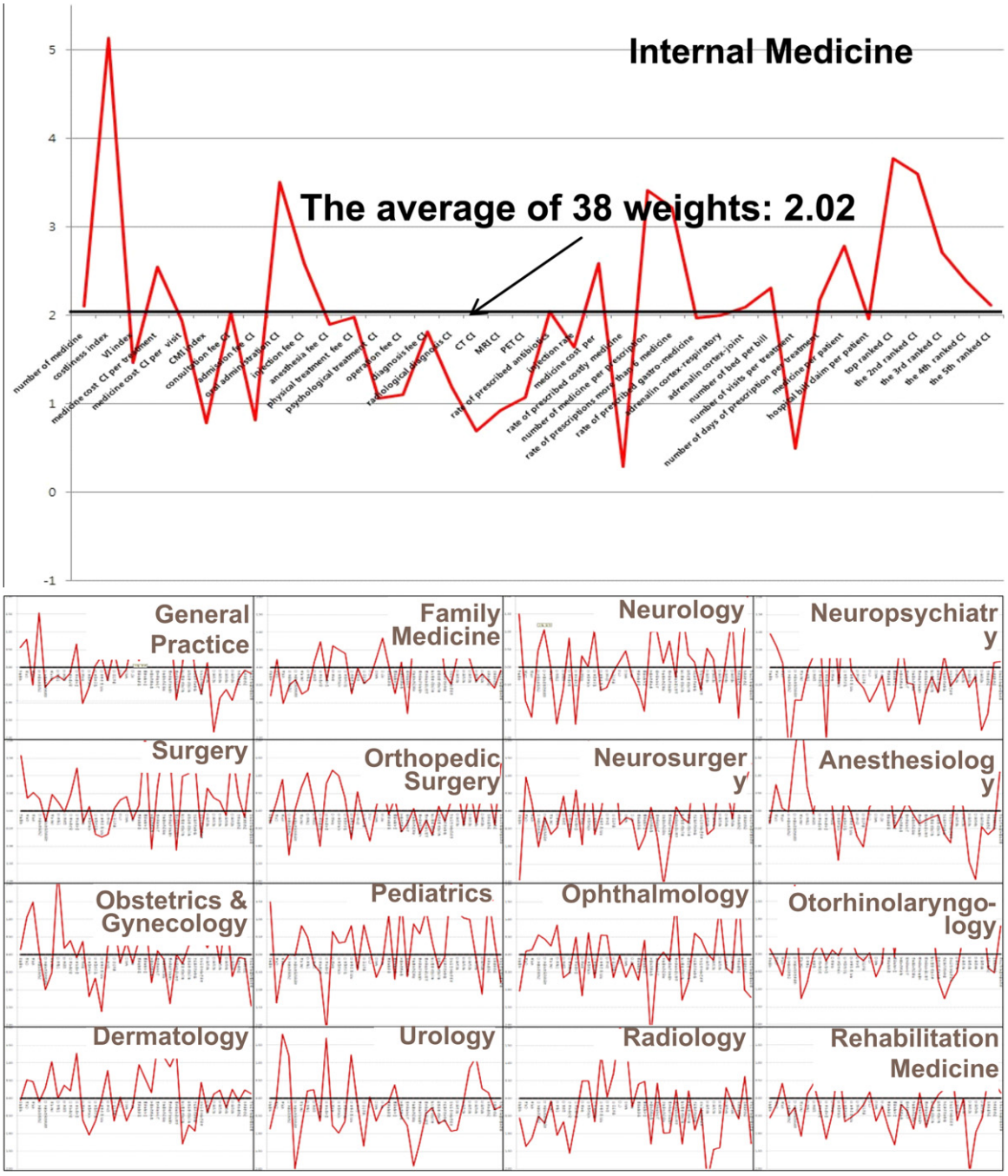

Fig. 3. Weights of 38 indicators used to compute the CDA index.

Table 1
Grade for degree anomalies for internal medicine clinics, $N = 3705$.

GDA	CDA	Log CDA	Frequency (%)
4	1000+	6.9+	236 (6)
3	100+	4.6+	154 (4)
2	10+	2.3+	570 (15)
1	5+	1.6+	403 (11)
0	~5	~1.00	2342 (63)

provides them with diagnostic information indicating the reasons a provider is selected for the intervention and helps determine the correction for the problematic behavior for a specific provider.

4. Validation and discussion

The mean of the log-transformed CDA for the non-intervention group is 2.23 whereas the mean for the intervention group is 3.97, and the likelihood ratio, p(CDA|intervention)/p(CDA|non-intervention), increases as the CDA score increases, which indicates that measuring providers' utilization patterns with the CDA score modeled in this study is consistent with the current manual selection process. Table 2 shows the two confusion matrices by the intervention history and the group defined by the CDA score — one with the cutoff at the top-10‰ of the CDA and the other with the cutoff at the top-30‰. We expected the proportion of providers in the intervention group to be similar in proportion to those under the

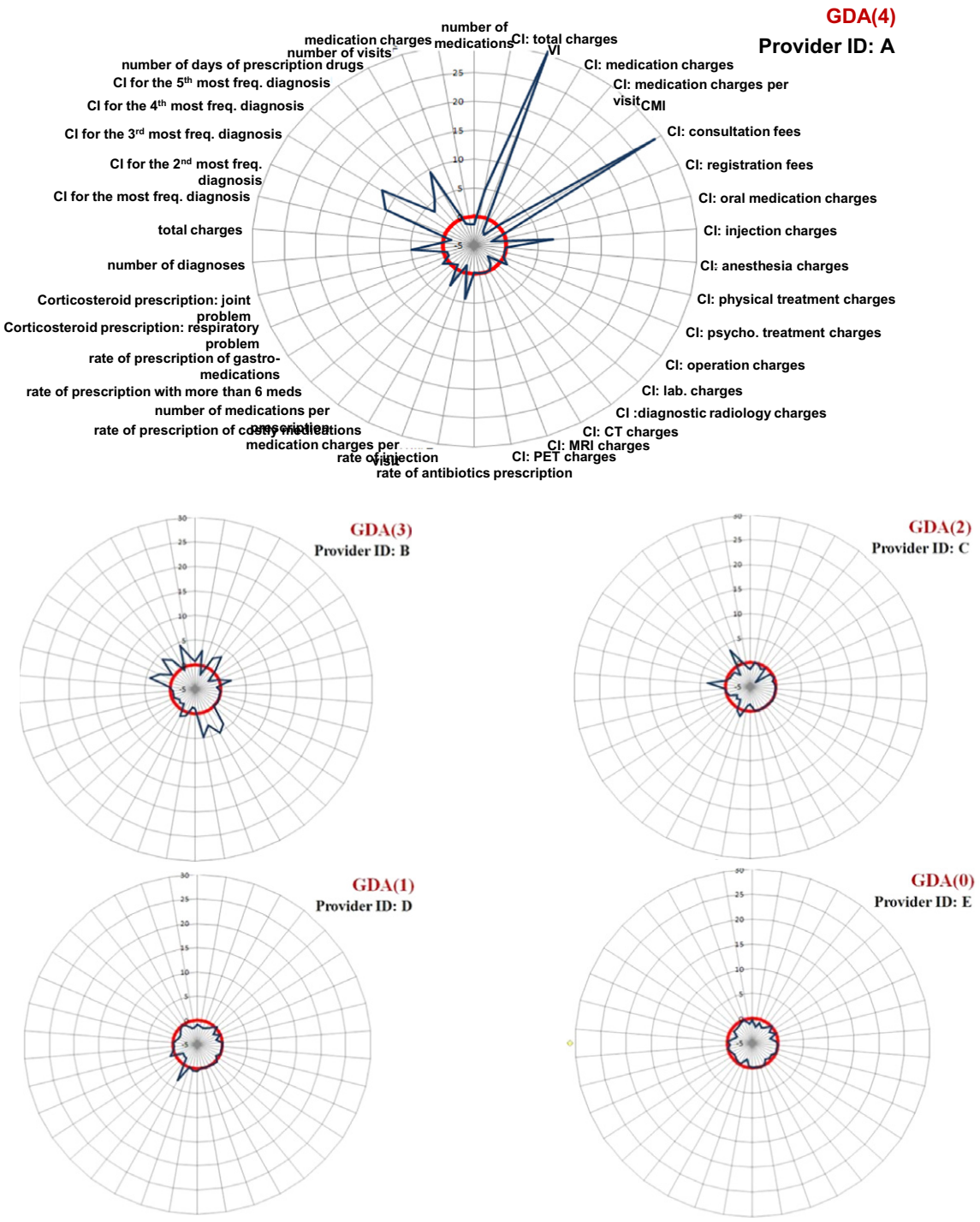

Fig. 4. Typical DA patterns of providers in each GDA of internal medicine.

manual review in the first matrix: 9.7%. The ratios of the average log-transformed CDA of the providers in cells (a) and (d), and the cells (e) and (h) are of the magnitude of 10, which implies the CDA score follows the manual review results well. The CDA score is reasonably consistent with the HIRA's practice of payment denial for claims it found inappropriate as well. The averages of percent of payment denied (PPD) for clinics in the cells (c), (d), (g), and (h), proposed intervention groups, are larger than the averages in the cells (a), (b), (e), and (f), proposed non-intervention groups, respectively. The averages for clinics in the cells (c) and (g), actual non-intervention groups, are larger than the averages in the cells (b) and (f), actual intervention groups, and the maximum PPD in the cells (c) and (g) was 20.4%. Obviously, HIRA reviewers missed the clinic whereas the proposed CDA model identified it as the one who needs intervention.

The cells (b), (c), (f), and (g) reveal inconsistencies between results from the manual review process and results from the proposed model. A couple of reasons account for the inconsistency. For solely educational purposes, the HIRA regularly selects providers with healthy and normal utilization behaviors for the intervention. It excludes providers from the selection that have undergone past intervention. We validated the proposed model against the HIRA's detection of abusive providers even though accuracy of the decision is in question. Unlike insurance fraud and abuse cases,

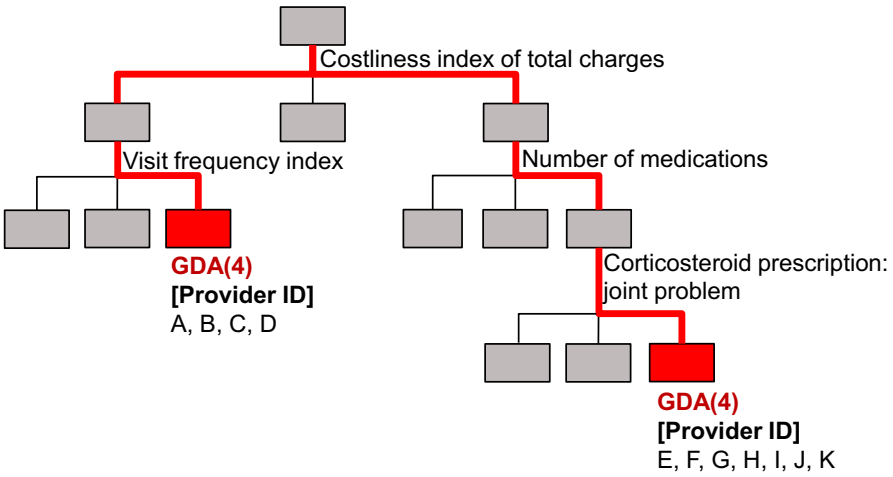

Fig. 5. Examples of segments of the GDA(4) in internal medicine and indicators that resulted in the segmentation.

Table 2
Confusion matrices by the intervention history and the group by the CDA score.

Intervention history			CDA < threshold$_{10\%}$		CDA ⩾ threshold$_{10\%}$		CDA < threshold$_{30\%}$		CDA ⩾ threshold$_{30\%}$
Non-intervention(N = 3346)	n	(a)	3044 (91%)	(c)	302 (9%)	(e)	2451 (73%)	(g)	895 (27%)
	Avg. Log(CDA)		1.18		12.77		0.75		4.28
	Avg. PPD[a]		0.30		0.45		0.28		0.40
	[Min Max]		[0.00 10.78]		[0.00 20.40]		[0.00 3.98]		[0.00 20.40]
Intervention(N = 359)	n	(b)	290 (81%)	(d)	69 (19%)	(f)	143 (40%)	(h)	216 (60%)
	Avg. Log(CDA)		2.17		11.53		1.20		7.80
	Avg. PPD[a]		0.38		0.48		0.38		0.41
	[Min Max]		[0.00 4.80]		[0.00 4.98]		[0.00 4.80]		[0.00 4.98]

[a] Percent of payment denied by the HIRA = 100 × (amount of payment denied)/(total charges claimed by a clinic).

the appropriateness of medical insurance claims, due to uncertainty behind the appropriate delivery of medical care, is difficult to ascertain (Henderson, 2009; Simborg, 2008). Reviews of every piece of utilization are prohibitively expensive. The inconsistency may imply imperfections in current practice. Manually reviewing indicators individually generates an enormous amount of data and the HIRA misses providers with serious problems. The average of the log-transformed CDA of the group in cell (c) (Table 2), providers not in the HIRA intervention group that would receive interventions based on the proposed model, raises the most serious question about the effectiveness of current HIRA practice.

Concerns abound over the accuracy of diagnosis coding in NHI claims in Korea (Park & et al., 2000; Shin et al., 1998). Providers have been accused of routinely up-coding and over-coding diagnoses to prevent HIRA denial of fee-for-service payment. However, the model proved to be robust against the reliability of diagnostic information by picking the indicators that use the diagnosis-based classification to adjust for differences in the case-mix of providers as importance variables.

5. Conclusion

In this study, we proposed a scoring model that measures the abusiveness of healthcare providers in the medical services claims submitted to a payer. The model is composed of two parts: scoring and segmentation. Through scoring, the model quantifies the degree of abusiveness, and based on these scores, the segmentation is used to group the problematic providers with similar profiles in the billing pattern. The significance of this study is it suggested a model that can be customized to detect abusive utilization patterns in various payment arrangements.

Three unique features characterize the proposed method. First, the scoring model alerts payers with information that an anomalous billing pattern, which may indicate abusive utilization behavior, has been found. Although they do not allow for certainty about abusive and normal billing patterns, the scores allow stakeholders to select providers that need to be examined more closely. Because of the expense in undertaking a detailed investigation of all claims submitted for insurance payment, a payer may concentrate investigation on only those thought most likely to be abusive. Second, while the scoring model integrates multiple attributes, it offers information on the attribute most dissimilar from the norm. This output differs from that of other machine learning or data mining algorithms, which tend to implicitly embed the mechanism of the detection procedure. Third, the segmentation model, based on decision trees, can be applied with greater efficiency to the resulting scores. It can be used to explain the reason a certain group of providers was chosen for further investigation and intervention, and payers can use the information in designing corrective measures for the group.

Bolton and Hand (2002) reported that types of fraud are growing increasingly sophisticated, and patterns detected from fraudulent and non-fraudulent behaviors quickly become obsolete because of the rapid changes in behavior. The proposed model is flexible, scalable and easy to use and update. A medical insurance payer can use it to design the selection rule that best utilizes claims reviewers and domain experts for thorough investigation of potential abusers while controlling expenses.

The major effort to contain healthcare costs during the past several decades has been focused on the bundling of payment units and changes in incentive structures of providers and other stakeholders. Our approach may present other opportunities for fighting

escalating healthcare costs. It is particularly relevant because more medical information is stored as electronic data with the diffusion of information technology in healthcare, particularly through penetration of electronic health records.

This work motivates possible future studies. The full application for each specialty requires continued refinement, although the method presented in this study is general. By broadening the number of specialty-specific attributes (or indicators) more can be selectively integrated together.

Acknowledgements

This work was conducted by a research grant from the Health Insurance and Review Agency (HIRA 2007-82), and the authors thank the HIRA staff for their invaluable assistant and comments. Dr. Shin's work on this project was partially supported by the research funding program, Post Brain Korea 21 and a research grant from the National Research Foundation of Korea (2009-0065043/2011-0018257). Dr. Jhee's work was partially supported by a research grant from Hongik University.

Appendix A

Definition of indicators

No.	Indicator	Definition
1	CI: total charges	Case-mix adjusted costliness index of total charges[a]
2	VI	Case-mix adjusted visit frequency index[a]
3	CI: medication charges	Case-mix adjusted costliness index of medication charges[a]
4	CI: medication charges per visit	Case-mix adjusted costliness index of medication charges per visit[a]
5	CMI	Case mix index[b]
6	CI: consultation fees	Case-mix adjusted costliness index of consultation fees[a]
7	CI: registration fees	Case-mix adjusted costliness index of registration fees[a]
8	CI: oral medication charges	Case-mix adjusted costliness index of charges for oral medication[a]
9	CI: injection charges	Case-mix adjusted costliness index of charges for IV injections[a]
10	CI: anesthesia charges	Case-mix adjusted costliness index of charges for anesthesia[a]
11	CI: physical therapy charges	Case-mix adjusted costliness index of charges for physical therapy[a]
12	CI: psychiatric treatment charges	Case-mix adjusted costliness index of charges for psychiatric treatments[a]
13	CI: operation charges	Case-mix adjusted costliness index of charges for operating room procedures[a]
14	CI: lab. charges	Case-mix adjusted costliness index of charges for laboratory tests[a]
15	CI: diagnostic radiology charges	Case-mix adjusted costliness index of charges for diagnostic radiology[a]
16	CI: CT charges	Case-mix adjusted costliness index of charges for computed tomography[a]
17	CI: MRI charges	Case-mix adjusted costliness index of charges for magnetic resonance imaging[a]
18	CI: PET charges	Case-mix adjusted costliness index of charges for positron emission tomography[a]
19	Rate of antibiotics prescription	(Number of claims with prescription of antibiotics)/(total number of claims)
20	Rate of injection	(Number of claims with prescription of IV injections)/(total number of claims)
21	Medication charges per visit	Average medication charges per visit
22	Rate of prescription of costly medications	(Number of prescriptions with costly medications monitored by the HIRA)/(total number of prescriptions)
23	Number of medications per prescription	Average number of medications per prescription
24	Rate of prescription with more than six medications	(Number of prescriptions with more than six medications)/(total number of prescriptions)
25	Rate of prescription with gastro-medications	(Number of prescriptions with gastro-medications)/(total number of prescriptions)
26	Corticosteroid prescription: respiratory problem	(Number of claims with respiratory principal diagnosis[c] and prescription of corticosteroids)/(total number of claims with respiratory principal diagnosis[c])
27	Corticosteroid prescription: joint problem	(Number of claims with principal or secondary diagnosis of joint problem[d] and prescription of corticosteroids)/(total number of claims with principal or secondary diagnosis of joint problem[d])
28	Number of diagnoses	Average number of diagnoses per claim
29	Total charges	Average total charges per claim
30	CI for the most freq. diagnosis	Case-mix adjusted costliness index of total charges for claims with the principal diagnosis that is the most frequent at the clinic[a]
31	CI for the 2nd most freq. diagnosis	Case-mix adjusted costliness index of total charges for claims with the principal diagnosis that is the second most frequent at the clinic[a]
32	CI for the 3rd most freq. diagnosis	Case-mix adjusted costliness index of total charges for claims with the principal diagnosis that is the third most frequent at the clinic[a]

(continued on next page)

Appendix A (*continued*)

No.	Indicator	Definition
33	CI for the 4th most freq. diagnosis	Case-mix adjusted costliness index of total charges for claims with the principal diagnosis that is the fourth most frequent at the clinic[a]
34	CI for the 5th most freq. diagnosis	Case-mix adjusted costliness index of total charges for claims with the principal diagnosis that is the fifth most frequent at the clinic[a]
35	Number of medication days	Average number of days prescribed with medication per claim
36	Number of visits	Average number of visits per claim
37	Medication charges	Average medication charges per claim
38	Number of medications	Average number of medications per claim

[a] $CI_{ij} = \frac{\sum_{k=1}^{K} n_{ik}\bar{x}_{ijk}}{\sum_{k=1}^{K} n_{ik}\bar{X}_{jk}}$ Eq. (A.1) where i = subscript for a clinic, j = subscript for a type of charge or utilization, k = subscript for a patient classification group (Korean Outpatient Group, KOPG), K = total number of KOPGs, n_{ik} = number of claims classified to the KOPG k submitted by the clinic i in the quarter, $\bar{x}_{ijk}$ = average *charge of* j for claims in the KOPG k submitted by the clinic i in *the quarter*, and $\bar{X}_{jk}$ = average charge of j for claims in the KOPG k submitted by all clinics in the *specialty in the previous year*.

[b] $CMI_i = \frac{\sum_{k=1}^{K} n_{ik}\bar{X}_k}{\bar{X}}$ Eq. (A.2) where i = subscript for a clinic, k = subscript for a KOPG, K = total number of KOPGs, n_{ik} = number of claims classified to the KOPG k submitted by the clinic i in the quarter, $\bar{X}_k$ = average total charge for claims in the patient group k submitted by all clinics in the specialty in the previous year, $\bar{X}$ = average total charge for all claims in the specialty submitted in the previous year.

[c] ICD-10 (the 10th revision of the International Classification of Diseases) principal diagnosis of J00∼J44, or J47.

[d] ICD-10 principal or secondary diagnosis of M13∼M17, or M19.

References

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*, 235–255.

Bonchi, F., Giannotti, F., Mainetto, G., & Pedreschi, D. (1999). A classification-based methodology for planning auditing strategies in fraud detection. In *Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining,* San Diego, CA, 15–18 August (pp. 175–184).

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York: Chapman and Hall.

Brockett, P., Xia, X., & Derrig, R. A. (1998). Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. *Journal of Risk and Insurance, 65*, 245–274.

Center for Medicare and Medicaid Services (2008a). Medicare claims review programs: MR, NCCI edits, MUEs, CERT, and RAC. Baltimore: CMS. http://www.cms.hhs.gov/MLNProducts/downloads/MCRP_Booklet.pdf. Accessed Jul 13, 2009.

Center for Medicare and Medicaid Services (2008b). The improper Medicare FFS payments report— May 2008. Baltimore: CMS. http://www.cms.hhs.gov/apps/er_report/preview_er_report_print.asp?from=public&which=long&reportID=9. Accessed July 14, 2009.

Chan, C., & Lan, C. (2001). A data mining technique combining fuzzy sets theory and Bayesian classifier—an application of auditing the health insurance fee. In *Proceedings of the international conference on artificial intelligence,* Las Vegas, NV, 25–28 June (pp. 402–408).

Cox, E. (1995). A fuzzy system for detecting anomalous behaviors in healthcare provider claims. In S. Goonatilake & P. Treleaven (Eds.), *Intelligent systems for finance and business* (pp. 114–134). New York: John Wiley and Sons.

Eisenberg, J. M. (2002). Physician utilization: The state of research about physicians' practice patterns. *Medical Care, 40*, 1016–1035.

Feldman, R. (2001). An economic explanation for fraud and abuse in public medical care programs. *Journal of Legal Studies, 30*, 569–577.

Guyon, I., & Elisseeff, A. (2003). Introduction to variable and feature selection. *Machine Learning, 3*, 1157–1182.

Hager, G., Upton, C., Graycarek, R., Knowles, V., McNees, E., & Perry, J. (2006). Information systems can help prevent, but not eliminate, health care fraud and abuse. Frankfort: Kentucky Legislative Research Commission. http://www.lrc.ky.gov/lrcpubs/RR%20333_forweb.pdf. Accessed July 14, 2009.

He, H., Wang, J., Grac, oW., & Hawkins, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications, 13*, 329–336.

He, H., Hawkins, S., Graco, W., & Yao, X. (2000). Application of genetic algorithm and k-nearest neighbour method in real world medical fraud detection problem. *Journal of Advanced Computational Intelligence and Intelligent Informatics, 4*, 130–137.

Health Insurance Review and Assessment Service & National Health Insurance Corporation (2008). 2007 National Health Insurance Statistical Yearbook. Seoul: Health Insurance Review & Assessment Service and National Health Insurance Corporation. http://www.hira.or.kr/common/dummy.jsp?pgmid=HIRAF010303000000. Accessed July 14, 2009.

Henderson, J. W. (2009). *Health economics & policy* (4th ed.). Mason, OH: South-Western Cengage Learning (pp. 86–89).

Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *Journal of Healthcare Information Management, 19*, 64–72.

Li, J., Huang, K. Y., Jin, J., & Shi, J. (2008). A survey on statistical methods for health care fraud detection. *Health Care Management Science, 11*, 275–287.

Liu, H., & Motoda, H. (2001). *Feature extraction, construction and selection: A data mining perspective*. Norwell: Kluwer Academic Publishers.

Major, J., & Riedinger, D. (2002). EFD: A hybrid knowledge/statistical-based system for the detection of fraud. *Journal of Risk and Insurance, 69*, 309–324.

Matignon, R. (2007). *Data mining using SAS enterprise miner*. Hoboken: Wiley-Interscience.

National Health Care Anti-Fraud Association (2009). The problem of health care fraud. National Health Care Anti-Fraud Association Web. http://www.nhcaa.org/eweb/DynamicPage.aspx?webcode=anti_fraud_resource_centr &wpscode=TheProblemOfHCFraud. Accessed July 14, 2009.

Organisation for Economic Co-operation and Development (2008). Health at a glance 2007: OECD indicators. Paris: OECD Publishing. http://www.sourceoecd.org/socialissues/9789264027329. Accessed Jul 13, 2009.

Organisation for Economic Co-operation and Development (2009). OECD health data 2009: Frequently requested data. [Internet]. OECD. http://www.oecd.org/document/16/0,3343,en_2649_34631_2085200_1_1_1_1,00.html. Accessed Jul 13, 2009.

Ortega, P. A., Figueroa, C. J., & Ruz, G. A. (2006). A medical claim fraud/abuse detection system based on data mining: A case study in Chile. In *Proceedings of the 2006 international conference on data mining,* Las Vegas, NV, (pp. 224–231).

Park, J. K., Kim, K. S., Kim, C. B., Lee, T. Y., Lee, K. S., Lee, D. H., et al. (2000). The accuracy of ICD codes for cerebrovascular diseases in medical insurance claims. *Journal of Preventive Medicine Public Health, 33*, 76–82.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2005). A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review, 23*, 1–14.

Pontell, H. N., Jesilow, P. D., & Geis, G. (1982). Policing physicians: Practitioner fraud and abuse in a government medical program. *Social Problems, 30*, 117–125.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco: Morgan Kaufmann Publishers Inc.

Rai, A. K. (2001). Health care fraud and abuse: A tale of behavior induced by payment structure. *Journal of Legal Studies, 30*, 579–587.

Shane, R. (2000). Detecting and preventing health care fraud and abuse—we've only just begun. *American Journal of Health System Pharmacy, 57*, 1078–1080.

Shin, E., Park, Y. M., Park, Y. G., Kim, B. S., Park, K., & Meng, K. (1998). Estimation of disease code accuracy of national medical insurance data and the related factors. *Journal of Preventive Medicine and Public Health, 31*, 471–480.

Simborg, D. W. (2008). Health fraud: Whose problem is it anyway? *Journal of the American Medical Information Association, 15*, 278–280.

US Federal Bureau of Investigation (2009). Financial crimes report to the public: Fiscal year 2007. Washington, DC: FBI. http://www.fbi.gov/publications/financial/fcs_report2007/financial_crime_2007.htm#health. Accessed July 14, 2009.

Viaene, S., Richard, A., & Dedene, D. (2005). A case study of applying boosting Naive Bayes to claim fraud diagnosis. *IEEE Transactions on Knowledge and Data Engineering, 16*, 612–620.

Yamanishi, K., Takeuchi, J., Williams, G., & Milne, P. (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery, 8*, 275–300.

Yang, W. S., & Hwang, S. Y. (2006). A process-mining framework for the detection of healthcare fraud and abuse. *Expert Systems with Applications, 3*, 56–58.