Identifying Doctors' Overcharge in Healthcare Claims Using Bayesian Network Model

Huanye Liu[*]

June 2017

Abstract

We describe a Bayesian network model based method for ranking and detecting candidate audit doctors from a data set of medicare claims. The relevant audit doctors are attending physicians who exhibit certain statistical behavior indicative of potential overcharging outpatients based on medicare claims during the period from 2008 to 2010. The approach used here is consistent with several previous works applying statistical methods to fraud and waste detection, but has its own emphasis on three specific aspects: first, based on relevant domain knowledge, several key features highly related to the overcharge behavior of attending physicians are selected from the outpatients medicare claims; second, a Bayesian network based statistical model is developed to characterize relationship among all key features and the target variable, the average medicare claim payment; and third, statistical hypothesis testing using dynamic programming is applied to identify attending physicians who diverge significantly from their expected behavior according to inference results of the Bayesian network model.

keywords: Bayesian Network Model, fraud and waste detection, overcharge behavior, dynamic programming.

[*] University of Chicago, MACSS, 5757 S. University Avenue Chicago, Illinois 60637, huanyeliu@uchicago.edu.

1 Introduction

Identifying fraud and abuse in healthcare claims, doctor's overcharge behavior in particular, previously relied heavily on specialized knowledge and forensic skills of human expertise, and thus was normally considered to be a human intensive job. However, the ever-growing number of healthcare claims now calls for the development of a suitable computer-aided methodology for fraud and abuse detection in the healthcare system, which requires extracting domain knowledge from data using computer-aided analysis to replace human expertise from professionals. The implementation of a computer-aided methodology is supposed to be based on various factors related to medical diagnoses, considering all kinds of procedures and treatment protocols and the subtleties of the prescribed medications. All included, these factors would form a high dimensional set of predictors, so we find an appropriate computational model, the Bayesian network, to store and process the claims data and to capture relationship among all these factors for further analysis in an efficient way.

The extensive use of the probabilistic inference methods such as the probabilistic expert systems [1], causal models [2], and Bayesian networks [3] in many other areas of healthcare also motivates us to adopt the Bayesian network to analyze the healthcare claims data. These areas include patient care management [4] [5], gene regulation networks [6], diagnosis systems [7], disease and infection [8] and bioinformatics and computational biology [9].

The basic classification of healthcare fraudulent behaviors is described in [10], which also briefly explains why we should use statistical methods for detecting fraud and abuse in various scenarios such as the phantom or duplicate billing, identity theft, medicare forgery, fictitious or deceased beneficiaries, bill padding, and medicare forgery [11]. But the method proposed in this paper for fraud and abuse detection in medicare claims data is to identify doctors who are associated with abnormal and excessive number for high claim payment. We do not attempt to develop statistical models for the underlying mechanism of any of these abnormal and excessive medcares here; we instead identify the deviations from normative or baseline behavior, considering all factors which characterize the interaction between the patient and the attending physician, and then aggregates over the set of claims for each attending physician.

Fraud detection in the healthcare system can be carried out in either the off-line or online modes of analysis [12]. The off-line mode can be used by audit investigators to retrospectively review claims data for identification of fraud or abuse for further investigation. The online mode instead puts emphasis on early detection of a potential fraudulence behavior in order to take actions as early as possible to prevent further abuse or fraud, and thus can help to eliminate unnecessary waste in the healthcare system in time.

The approach we take here has the same nature as other previous works [12] [13] [14] on statistical methods for detection of fraud and abuse in various domains such as telecommunications, financial trading, network intrusion, health care and credit card transactions. However, our approach is more of an unsupervised statistical method which does not require labeling fraudulent claims explicitly. In contrast, the supervised or semi-supervised statistical methods may directly model the fraud outcomes in terms of all other related predictors. Due to the ever-growing number of healthcare claims and rapidly-changing nature of fraud and abuse incidences, the labeling of fraudulent claims becomes more and more difficult, which justifies the approach we propose here.

Two previous works [12] [15] also do not require any explicit labeling of fraudulent claims in the analysis, and here we can compare each of them with the method proposed in this paper. For example, [12] used both the off-line and online application modes for medicare fraud detection. For each drug, the previous claims data were used to obtain pairwise occurrence frequencies in the individual medicare combined with other factors including age, gender, medical diagnosis, or other prescribed medications. For each of these co-occurrence dimensions, the likelihood of fraud was associated with the particular medication in a given medicare claim. This likelihood was assigned a very high value when compared with appropriate reference frequencies while the other factors in the medicare claim of interest small values of the relevant occurrence frequencies. Based on domain expertise, some appropriate thresholds were set for each likelihood score, and claims with likelihood scores above these thresholds were classified as fraudulent claim.

Our approach, compared with the approach in [12], is to identify attending physicians who exhibit overcharge behavior in the offline mode based on multiple claims instances which can be used to model the normal or expected medicare behavior, and this baseline model for normal behavior should capture the more complex and high-dimensional interactions between healthcare providers and patients recorded in claims data. For example, the medical history of a patient including patient medication profiles which are collected from the anonymized claims data, can play the role of training data for the normalizations of claim charge behavior and predicting the expectation of claim charge behavior, and this kind of training data should also include attending physician attributes such as their pattern of diagnoses and procedures. The reason why we need a multivariate behavioral models for the analysis of sparse, high-dimensional data has been mentioned before with several examples.

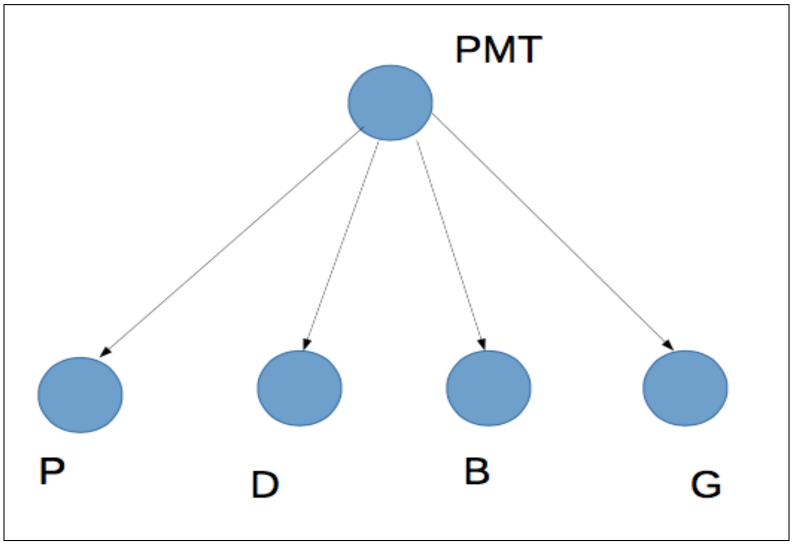
The paper [15] is mainly concerned with expense auditing for corporate travel and entertainment, and we adopt very similar approach and methodology proposed in [15]. For example, [15] examined expense claims submitted by employees in various focus areas including ground transportation, restaurant tip, etc., and evaluated the expense claims of different entities such as individuals or entire departments in these areas. A normalized baseline model for each focus area was defined, and for each entity, any abnormal behavior was detected if there were significant departures from the expected behavior. This approach assumes that any abnormal behavior only takes a small fraction of the overall set of expense claims so that normalized baseline models can be reliably estimated. [15] used the likelihood ratio score to identify the entities with abnormal behavior, which can be computed as the likelihood ratio of the actual behavior over the set of all related variables for each entity to the predictions of the normalized baseline model for this same set of related variables. [15] also used Monte Carlo methods, which are similar to those used in scan statistics [16] to evaluate the relevant p-values of the likelihood ratios. Although the general methods proposed in [15] and in our paper are similar, applying that method to the healthcare medicare fraud is significantly more challenging due to the complexity of the data and the related model. Therefore, we adopt the graphical model, the Bayesian network model in particular, to accommodate the sparse, high-dimensional data.

Our research question is how to develop Bayesian network model based methods to detect doctors' overcharge behavior. The motivation come from the facts that doctors' overcharge behavior is still rampant, which incurs a large amount of cost/waste in the healthcare system each year. Therefore, to prevent such kind of misbehavior in the healthcare system, we need to devise effective computer aided methods to measure and detect the overcharge behavior, and based on the results, we then take appropriate actions to restrict doctors' overcharge behavior and thus reduce the unnecessary waste in the healthcare system.

2 Model

We use Bayesian network model to build the multivariate relationship among key variables in this project. To detect doctors' overcharge behavior, we need to first operationalize the concept of overcharge. In this project, it is defined as the relatively large number of highest claim payment submitted by one doctor compared with the normal number or expected number of the highest claim payment the doctor should have submitted. The Bayesian network is shown as below:

Figure 1: Bayesian Network Model



First of all, we compute the posterior probability of the number of highest payment NPI conditioning on all other variables which include claim procedure code P, claim diagnostic code D, patient's year of birth B and patient's gender G. This posterior probability measures normally how likely a doctor submits such number of highest claim payments given the patient's age, the patient's gender, the patient's disease type and the patient's medical intervention, and it can be estimated from Bayes' rule:

$$Pr(PMT|P,D,B,G) = \frac{Pr(PMT,P,D,B,G)}{\sum_{PMT} Pr(PMT,P,D,B,G)} \tag{1}$$

The joint probability $Pr(PMT,P,D,B,G)$ can be further split into five components thanks to the Bayesian network and structure:

$$Pr(PMT,P,D,B,G) = Pr(PMT)Pr(P|PMT)Pr(D|PMT)Pr(B|PMT)Pr(G|PMT) \tag{2}$$

where

$$\begin{aligned}
Pr(P|PMT) &= \frac{N(P, PMT)}{\sum_P N(P, PMT)} \\
Pr(D|PMT) &= \frac{N(D, PMT)}{\sum_D N(D, PMT)} \\
Pr(B|PMT) &= \frac{N(B, PMT)}{\sum_B N(B, PMT)} \\
Pr(G|PMT) &= \frac{N(G, PMT)}{\sum_G N(G, PMT)}
\end{aligned} \tag{3}$$

So what we need to do is to count for each kind of claim payment PMT, for each type of X, how many claims are there including the pair (X, PMT), where X is over P, D, B and G. Since the medicare claim data we use has relatively big size, we choose the MapReduce library in python as the computational tool to implement the task.

Given all estimated posterior probabilities from above, the next step we use dynamic programming to (a) accumulate these probability across one particular doctor to identify whether the doctor has submitted too large number of highest payments conditioning on all other variables, and (b) we also compute the expected or normal number of all types of payments that the doctor has submitted conditioning on all other variables, and compare the histogram of expected number of all types of payments with the histogram of actual number of all types of payments submitted by the doctor, and use the chi-square statistics to measure the difference between the two histograms for abnormal charge behavior identification.

For step (a), we need to figure out how likely the expected number of highest claim payments is above the observed number of highest claim payments submitted by that particular doctor:

$$Pr(N_{pmt} >= O_{pmt}) = \sum_{k=o_{pmt}}^{T} Pr(P_{pmt} = k) \tag{4}$$

where T is the total number of all types of claim payments submitted by the doctor and the O_{pmt} is the observed number of highest claim payments and each $Pr(P_{pmt} = k)$ is the probability that the doctor has submitted k highest claim payments among all T claim payments. Now introduce the notation $f_{k,i}$ which is the probability that the doctor has submitted k highest claim payments among all i claim payments, we need to figure out $f_{k,T}$ for all $k = i, ..., T$ and sum them up to compute $Pr(N_{pmt} >= O_{pmt})$. Using dynamic programming, we have the following recursive relation with the initial values $f_{0,0} = 1$ and $f_{k,0} = 0$ for all $k = i, ..., T$:

$$\begin{aligned} f_{k,i} =& Pr(the \quad i_{th} \quad payment \quad is \quad the \quad highest \quad payment|y_i)f_{k-1,i-1} \\ &+ Pr(the \quad i_{th} \quad payment \quad is \quad not \quad the \quad highest \quad payment|y_i)f_{k,i-1} \end{aligned} \tag{5}$$

where y_i is all the categorical values of P, D, B, G of the i_{th} claim submitted by the doctor, so $Pr(the \quad i_{th} \quad payment \quad is \quad the \quad highest \quad payment|y_i)$ and $Pr(the \quad i_{th} \quad payment \quad is \quad not \quad the \quad highest \quad payment|y_i)$ are the posterior probabilities estimated previously from Bayes' rule.

For step (b), once we get all values of $f_{k,i}$ for $k = i, ..., T$ and $i = k, ..., T$, we can compute the expected number of all types of payments:

$$E_{pmt} = \sum_{k=1}^{T} Pr(N_{pmt} = k)k \tag{6}$$

for all payment types the doctor has submitted that pmt can take, and then we compare the histogram of expected number of all types of payments with the histogram of actual number of all types of payments using chi-square statistics to measure the doctor's abnormal charge behavior:

$$\chi^2 = \sum_{k=1}^{k} \frac{(O_k - E_k)^2}{E_k} \tag{7}$$

If the p-value from the chi-square statistic is too small, then we find a suspicious doctor with abnormal charge behavior.

3 Data

The data we use for this project, the CMS linkable 2008-2010 Medicare Data Entrepreneurs' Synthetic Public Use File, can be downloaded from the Centers for Medicare and Medicaid Services website: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/index.html. These Medicare Claims Synthetic Public Use Files was designed to create a new type of file that would be useful for data entrepreneurs, for software and application development, and for research training purposes.

The Data Entrepreneurs' Synthetic Public Use File contains multiple files per year for multiple years. The file contains synthesized data taken from a 5% random sample of Medicare beneficiaries in 2008 and their claims from 2008 to 2010. Each synthetic beneficiary was assigned a unique unidentifiable ID, which is provided on each file to link synthetic claims to a synthetic beneficiary. This beneficiary ID carries no information about the enrollee or any patient records, and is provided solely for reference and data processing purposes. The CMS Linkable 2008-2010 Medicare DE-SynPUF contains five types of files: the CMS Beneficiary Summary DE-SynPUF, the CMS Inpatient Claims DE-SynPUF, the CMS Outpatient Claims DE-SynPUF, the CMS Carrier Claims DE-SynPUF, and the CMS Prescription Drug Events (PDE) DE-SynPUF from 2008 to 2010. Table 1 describes the number of variables, the unit of record, and the number of records available in the full DE-SynPUF. Table 2 provides the number of observations available in each of the data files for each of the DE-SynPUF subsamples.

In this project, we choose the the CMS Outpatient Claims DE-SynPUF data sets after merging with the CMS Beneficiary Summary DE-SynPUF to include each patient's age and gender besides the procedure code, the diagnostic code and the claim payment information. The original CMS Outpatient Claims DE-SynPUF data set contains 76 variables and each record each record pertains to a synthetic outpatient claim. The original CMS Beneficiary Summary DE-SynPUF data set contains 32 variables and record pertains to a synthetic Medicare beneficiary. To answer our research question, we finally select a small portion of those variables and after all the data cleaning and munging, our data set of 161898 medicare claims includes 5 categorical variables NPI, P, D, B, G and the target variable PMT, the type of claim payment:

Table 1: Variables of Interest

Variable	Description	Number of Levels	Mode
NPI	Doctor's Unique Identification	161898	NA
P	Claim Procedure Code	3339	36415
D	Claim Diagnostic Code	12320	4019
B	Patient's Year of Birth	75	1940
G	Patient's Gender	2	2
PMT	Claim Average Payment	42	100

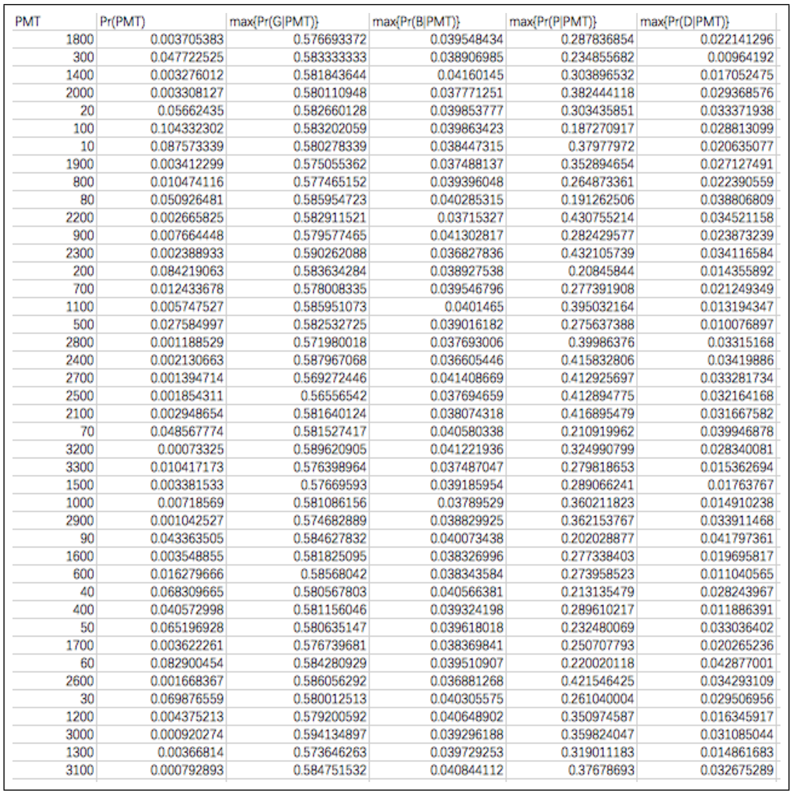
4 Estimation

Using the Bayesian network model specified in the Model Section, we estimate values for all the posterior probabilities. As we can see from the data section, there are in fact hundreds of thousands of possible pairs or combinations between different values of PMT and different values of P, D, B and G respectively even if we use Bayesian network model to avoid computing the joint probability directly. Given such large number of parameter estimates, the table below only shows the probability of each type of PMT $Pr(PMT)$ and the maximum conditional probability of one gender over both genders given each type of payment $\max_G Pr(G|PMT)$, the maximum conditional probability of one birth year over all birth years given each type of payment $\max_B Pr(B|PMT)$, the maximum conditional probability of one procedure code over all procedure codes $\max_P Pr(P|PMT)$and the maximum conditional probability of one diagnostic code over all diagnostic codes $\max_D Pr(D|PMT)$:

5 Experiment

Given all the estimated parameters, we apply dynamic programming to accumulate posterior probabilities across a doctor to identify whether the doctor has submitted a too large number of highest claim payments. we compute the probability that the normal number of highest claim payments is higher than the observed number of

Figure 2: Table of Estimated Parameters



PMT	Pr(PMT)	max{Pr(G|PMT)}	max{Pr(B|PMT)}	max{Pr(P|PMT)}	max{Pr(D|PMT)}
1800	0.003705383	0.576693372	0.039548434	0.287836854	0.022141296
300	0.047722525	0.583333333	0.038906985	0.234855682	0.00964192
1400	0.003276012	0.581843644	0.04160145	0.303896532	0.017052475
2000	0.003308127	0.580110948	0.037771251	0.382444118	0.029368576
20	0.05662435	0.582660128	0.039853777	0.303435851	0.033371938
100	0.104332302	0.583202059	0.039863423	0.187270917	0.028813099
10	0.087573339	0.580278339	0.038447315	0.37977972	0.020635077
1900	0.003412299	0.575055362	0.037488137	0.352894654	0.027127491
800	0.010474116	0.577465152	0.039396048	0.264873361	0.022390559
80	0.050926481	0.585954723	0.040285315	0.191262506	0.038806809
2200	0.002665825	0.582911521	0.03715327	0.430755214	0.034521158
900	0.007664448	0.579577465	0.041302817	0.282429577	0.023873239
2300	0.002388933	0.590262088	0.036827836	0.432105739	0.034116584
200	0.084219063	0.583634284	0.038927538	0.20845844	0.014355892
700	0.012433678	0.578008335	0.039546796	0.277391908	0.021249349
1100	0.005747527	0.585951073	0.0401465	0.395032164	0.013194347
500	0.027584997	0.582532725	0.039016182	0.275637388	0.010076897
2800	0.001188529	0.571980018	0.037693006	0.39986376	0.03315168
2400	0.002130663	0.587967068	0.036605446	0.415832806	0.03419886
2700	0.001394714	0.569272446	0.041408669	0.412925697	0.033281734
2500	0.001854311	0.56556542	0.037694659	0.412894775	0.032164168
2100	0.002948654	0.581640124	0.038074318	0.416895479	0.031667582
70	0.048567774	0.581527417	0.040580338	0.210919962	0.039946878
3200	0.00073325	0.589620905	0.041221936	0.324990799	0.028340081
3300	0.010417173	0.576398964	0.037487047	0.279818653	0.015362694
1500	0.003381533	0.57669593	0.039185954	0.289066241	0.01763767
1000	0.00718569	0.581086156	0.03789529	0.360211823	0.014910238
2900	0.001042527	0.574682889	0.038829925	0.362153767	0.033911468
90	0.043363505	0.584627832	0.040073438	0.202028877	0.041797361
1600	0.003548855	0.581825095	0.038326996	0.277338403	0.019695817
600	0.016279666	0.58568042	0.038343584	0.273958523	0.011040565
40	0.068309665	0.580567803	0.040566381	0.213135479	0.028243967
400	0.040572998	0.581156046	0.039324198	0.289610217	0.011886391
50	0.065196928	0.580635147	0.039618018	0.232480069	0.033036402
1700	0.003622261	0.576739681	0.038369841	0.250707793	0.020265236
60	0.082900454	0.584280929	0.039510907	0.220020118	0.042877001
2600	0.001668367	0.586056292	0.036881268	0.421546425	0.034293109
30	0.069876559	0.580012513	0.040305575	0.261040004	0.029506956
1200	0.004375213	0.579200592	0.040648902	0.350974587	0.016345917
3000	0.000920274	0.594134897	0.039296188	0.359824047	0.031085044
1300	0.00366814	0.573646263	0.039729253	0.319011183	0.014861683
3100	0.000792893	0.584751532	0.040844112	0.37678693	0.032675289

highest claim payments actually submitted by the doctor. Based on this probability, we rank doctors from those who are most likely overcharge patients to those who are least likely overcharge patients as this probability goes from small to large:

Table 2: Rank of Doctors Who Overcharge Patients

NPI	$Pr(N_{pmt} >= O_{pmt})$	Number of Highest Payments	Expected Number of Highest Payments
8687459280	0.0000051545	1	0.0000051545
4257804925	0.0000107343	2	0.0238377507
1728991257	0.0000107579	4	0.2830734919
8234450508	0.0000141459	2	0.0078447315
9863658854	0.0000239282	2	0.0121832158
589759662	0.0000300942	1	0.0000300942
2371152333	0.0000324308	1	0.0000324308
5810836174	0.0000327865	3	0.2343714115
7811358993	0.0000344331	2	0.0318564656
6006804581	0.0000363304	1	0.0000363304

As we can see from the table, if we compare the third column and the fourth column, we find the actual and expected numbers of highest claim payments differ a lot for these top ten doctors who are most likely overcharge. The second column, the probability that the normal number of highest claim payments is higher than the observed number of highest claim payments, is actually another mathematical equivalence of the difference between the third and the fourth column.

We can also check those doctors who have submitted large number of highest payments:

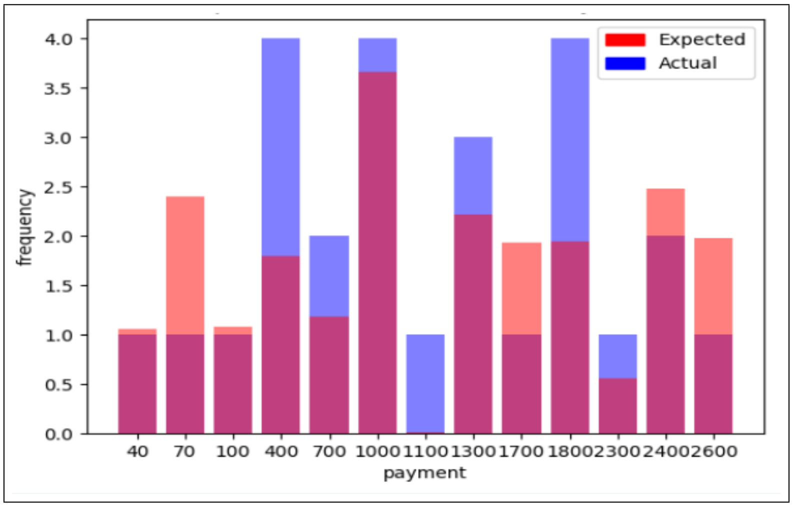
Table 3: Doctors Who Submitted Large Number of Highest Payments

NPI	$Pr(N_{pmt} >= O_{pmt})$	Number of Highest Payments	Expected Number of Highest Payments
3204358873	0.0480109606	14	8.9641158024
3733829210	0.7230098962	14	15.6702726184
8461982462	0.4153129262	14	12.9254407334
9943567290	0.3493275149	15	13.3559956746
9127235965	0.4230077793	16	14.9529958135
1308731628	0.4191260768	17	15.9035712276
2484358033	0.0703081096	19	13.5079929412
1586461282	0.865061931	24	28.805738982
2150853459	0.0349447214	24	16.6561038808
3177196045	0.1719397817	26	21.6349473007

Comparing the Table 3 with Table 2, we find that there is no overlap between these two tables, which means doctors who have submitted relatively large number of highest claim payments are not necessarily overcharge patients using our proposed methods; maybe they submitted large absolute number of highest claim payments just because their patients' disease are more sever and hence more costly. Therefore, our proposed model and methods incorporate all factors that is highly related to the claim payments and the thus would produce more convincing results that simply check the absolute number of highest payments submitted by the doctor.

Finally, we also compare the overall distribution of a doctor's all types of claim payments with the normal or expected distribution of those types of claim payments by using the chi-square test. The graph below shows the difference between these two distribution for a doctor who has a very low chi-square statistic p-value and therefore very likely has abnormal charge behavior:

Figure 3: Histogram Comparison with a very low χ^2 p-value



6 Conclusion

To sum up, we propose a methodology for the off-line application of overcharge detection using medicare claims data. The proposed approach uses Bayesian network to incorporate all related variables including diagnosis codes, procedure codes, medication history, and other doctor and patient attributes which characterize encounters between the doctor and the patient. The baseline model that we obtain captures the relationship among all above attributes under the Bayesian network framework, and any anomalous overcharge behavior would be identified based on the small p-value inferred from the the baseline model using computational aid of dynamic programming.

References

[1] D. J. Spiegelhalter, Probabilistic Expert Systems in Medicine: Practical Issues in Handling Uncertainty, Statistical Science, 2(1), pp. 25-30, 1987.

[2] J. Pearl, Causality: Models, Reasoning and Inference, 29, Cambridge: MIT press, 2000.

[3] F. V. Jensen and T. D. Nielsen, Bayesian networks and decision graphs Springer, 2007.

[4] P. J. F. Lucas, L. C. van der Gaag, A. Abu-Hanna, Bayesian networks in biomedicine and health-care, Artificial Intelligence in medicine, 30(3), pp. 201-214, 2004.

[5] M. Verduijn, N. Peek, P. M. J. Rosseel, E. de Jonge, B. A. J. M. de Mol, Prognostic Bayesian Networks: I: Rationale, Learning Procedure, and Clinical Use, Journal of Biomedical Informatics, 40(6), pp. 609-618, 2007.

[6] J. Donkers, K. Tuyls Computational Intelligence in Bioinformatics, Studies in Computational Intelligence, Springer, pp. 75-111, 2008.

[7] A.Meloni, A.Ripoli, V.Positano, L.Landini, Improved Learning of Bayesian Networks in Biomedicine, ISDA ?09, pp. 624-628, 2009.

[8] G. F. Cooper, D. H. Dash, J. D. Levander, W. K. Wong, W. R. Hogan and M. M. Wagner, Bayesian biosurveillance of disease outbreaks, Proceedings of the 20th conference on Uncertainty in artificial intelligence, pp. 94-103. AUAI Press, 2004.

[9] D. J. Wilkinson Bayesian methods in bioinformatics and computational systems biology, Briefings in Bioinformatics 8(2), pp. 109-116, 2007.

[10] Travaille P, Muller RM, Thornton D, van Hillegersberg J (2011) Electronic fraud detection in the U.S. medicaid healthcare program: lessons learned from other industries. In: Proceedings of the seventeenth americas conference on information systems.

[11] United States General Accounting Office (2000) Health care fraud. schemes to defraud, medicare, medicaid and private health insurers. GAO/T-OSI-00-15

[12] Aral KD, Guvenir HA, Sabuncuoglu I, Akar AR (2012) A prescription fraud detection model. Comput Methods Prog Biomed 106:37-46

[13] Hand DJ (2010) Fraud detection in telecommunications and banking: discussion of Becker, Volinsky and Wilks (2010) and Sudjianto et al. (2010), 52(1), 34-38

[14] Sudjianto A, Nair S, Yuan M, Zhang A, Kern D, Cela-Diaz F (2010) Stat Methods Fighting Financ Crimes 52(1):5-19

[15] Iyengar V, Boier I, Kelley K, Curatolo R (2007) Analytics for audit and business controls in corporate travel and entertainment. In: Proceedings sixth australasian data mining conference (AusDM 2007) CRPIT, vol 70. Gold Coast, pp 3-12.

[16] Kulldorff M (1997) A spatial scan statistic. Commun Statist Theor Meth 26(6):1481-1496