Expert Systems with Applications 37 (2010) 8628–8633

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Two models to investigate Medicare fraud within unsupervised databases

Rasim Muzaffer Musal *

Texas State University, 404 Rio Grande apt 209 Austin, TX 78701, United States

ARTICLE INFO

Keywords:
Fraud
Medicare
Unsupervised methods
Distances analysis
Clustering methods

ABSTRACT

We propose two models to identify fraud, waste and abuse in Medicare. These models are used to flag health care providers. The motivation for these models is based on observed cases of fraud. The paper details the use of clustering algorithms, regression analysis, and various descriptive statistics that are components of these models. Some of the challenges in the struggle to reduce fraud in Medicare are discussed.

1. Introduction

It is clear that fraud plays an important role in the US healthcare system. It increases the cost of healthcare through direct and indirect means. Its direct impact includes fraudulent monetary charges to the healthcare system. It also has indirect impacts which mainly arise from false positive identifications of fraudulent health care providers. These include opportunity costs associated with the medical education of the fraudulent providers and costs associated with the construction of complicated policies that effect beneficiaries and providers alike.

The general goal of organizations charged with fighting fraud can be formulated as reducing direct and indirect costs by maximizing the percentage of correct identification of fraudulent providers while minimizing the false ones with the least amount of resources. This paper exposes the application of two methodologies to identify fraudulent infusion therapy drug providers in a number of states from an unsupervised database.

The models employed in fraud identification can be generalized into two categories. There are models that identify fraud after observing nearly irrefutable amount of evidence, for instance identifying providers that charge for services rendered amounting to more than 24 h a day. On the other hand there are models which can identify providers of interest as being possibly involved in fraud after observing abnormal patterns in the data. The false positive rates for these models are likely to be higher. The models exposed in this paper belong to the latter category where we argue for their necessity in decreasing the overall level of fraud in the healthcare system.

The paper is organized in six sections. Section 2 is an overview of healthcare fraud investigations within US and provides a review of recent work in the literature but is not meant to be a comprehensive literature review. Section 3 provides an overview of data and Section 4 provides the exposition of the two models that we use to investigate fraud. In Section 5 we demonstrate the application of the methods. Section 6 includes the results and conclusion.

2. Overview

Anderson and Hussey (2001) present a bleak picture of the US healthcare system's performance when they compare the health care systems of 29 OECD countries. The comparison is made through various healthcare indicators among these countries. The authors find that for many of these indicators between the years 1960, 1980 and 1998 US' relative performance either did not improve or decline. The authors point out that US spends more than twice the median expenditure per capita among the OECD countries. They point out the larger GDP per capita of US relative to other OECD countries as a major component of this difference. However the authors also note that even when the GDP per capita is taken into account, there is still a large difference between observed and expected US healthcare expenditures. The authors' conclusion does not change when expected health care expenditures are calculated from the benchmark value of Switzerland, second ranked country in terms of health care expenditures per capita.

Anderson, Frogner, and Reinhardt (2007) investigate the health expenditures in the OECD countries. The authors are consistent in their findings with Anderson and Hussey (2001) and report that US spends more than 2.5 times per capita than the median country included in the study. Furthermore the authors report that US had fewer physicians, nurses, hospital beds, doctor visits and hospital days, per capita than the median OECD country. The authors' conclude that there are two major reasons why per capita spending is so much higher in USA. They suggest higher GDP per capita and higher prices account for these measures.

Part of the problem of higher prices lays in costs to the government programs arising from Fraud, Waste and Abuse (FWA). Li,

* Tel.: +1 202 625 1124.
E-mail address: mmusal@gmail.com

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.06.095

Huang, Jin, and Shi (2008) cite two different impact estimates. National Healthcare Anti Fraud association gave a conservative 3% estimate (60 Billion) of annual healthcare expenditure. FBI also reports the higher level of fraud at 10% (170 Billion).

By definition fraud requires intention on the part of the provider. Unfortunately we do not have any method to identify intention from the data. Therefore the methods outlined in this paper will also include in its goal to flag providers involved in waste and abuse. We use the term fraud below, instead of fraud waste and abuse for convenience.

Efforts to reduce fraud are complicated by several factors. There were more than 40 million beneficiaries that were eligible for Medicare part A and part B programs (Hoffman, Klees, & Curtis, 2008). Each year, these beneficiaries utilize services which result in terabytes of data. In other words, manual inspection of medical records is practically impossible. Heuristics that detect fraud such as duplicate claims within a database are useful but insufficient to catch but the most simplistic fraud. Furthermore, beneficiaries and health care providers are not homogenous. What constitutes as *normal* utilization rates and payments per claim of an oncologist is different from the *normal* utilization rates and payments of a family practitioner. The healthcare needs of a beneficiary will be dependent on his/her diagnosis and the provider's assessment of his/her individual needs. There are a large quantity of possible diagnosis and assessments. These factors lead to a large non fraud related variance in utilization and money amount paid per beneficiary. This adds to the challenge of, detecting abnormal patterns in data, which is the prevalent approach in detecting fraud within unsupervised databases.

Additional complications arise due to the dynamic nature of the system associated with healthcare fraud. Among the main components of this system are those who are charged with controlling and those who are agents of, fraud. In 1996, The Health Insurance Portability and Accountability Act (HIPAA) created the Health Care Fraud and Abuse Control Program (HCFAC). This program, as stated by the Department of Health and Human Services And The Department of Justice Health Care Fraud and Abuse Control Program Annual Report (2007) is "... designed to coordinate federal, state and local law enforcement activities with respect to health care fraud and abuse." The prevailing approach in the program emphasizes the monetary amount "recovered" as a result of investigation and targets those providers for whom an "extreme level" of evidence have been gathered. For instance the 2007 annual health care fraud and abuse control program report has two sections in its executive summary section. These are listed as monetary results and enforcement actions. In the monetary results section, the dollar amount earned for the federal government as well as the amounts transferred to the Medicare and Medicaid programs is listed. The monetary amounts listed involve the fiscal year 2007 as well as the total amount earned since 1997, the beginning of the HCFAC program. The enforcement actions list the number of cases that have been opened in 2007 as well as the number of defendants involved in these cases. The choice of the words in the titles for these sections as well as the lack of historical perspective in the enforcement actions section is implicative of the mindset in controlling fraud. The budget for the federal government in 2004 includes a section on performances of various programs (Performance and Management Assessments Budget of the United States Government Fiscal Year, 2004). The report states for the HCFAC program, "While providing some information on the status of fraud and abuse activities, the existing goals – return on investment, expected recoveries, and program savings – do not objectively measure if the program achieves its mission. The current measures do not demonstrate whether health care fraud and abuse have decreased, which is the program's ultimate mission." The Whitehouse archives contain a document on HCFAC "Detailed Information on the Health Care Fraud and Abuse Control Assessment" (2002) that refers to ongoing improvement plans which include quantitative measures of performance. However these measures do not address the issue raised by the 2004 report mentioned above.

As implied by Bolton and Hand (2002), Li et al. (2008) the prevalent methodology in the detection of fraud within unsupervised databases involves detecting possible outliers from claim amounts or utilization rates. We can classify providers involved in fraud in two broad categories of those who are high profile and those who are not. We use the term high profile providers as those providers who claim services to such an extent as to draw themselves attention in simple outlier models. The majority of the models employed in unsupervised databases target high profile providers which are relatively easier to identify. These models, in general have high true positive rates. For instance the *impossible days* model identify providers who charged for services that lasted longer than 24 h in a single day. The news release by the United State Attorney's Office District of Rhode Island (2008) indicates the successful application of this model. As useful as these models that target high profile providers are, if they are the only ones that are employed, they will not lower overall extent of fraud in Medicare. If these models are the only ones that are employed in fighting fraud, it would be unlikely to identify providers involved in fraud who take a more cautious approach. Even though the type of models this paper exposes would naturally have a higher false positive rate they can identify providers from the class of providers who are not high profile.

We use the term *behavioral* to describe the models in this paper as they are constructed to identify fraud based observed fraud cases. There are two major reasons why fraud investigation within unsupervised databases does not usually involve the type of models we expose in the paper. First of all, even though there is no formal comparison, it is likely that behavioral models flag more false positives than outliers of utilization rates and this increases costs to the healthcare system. Second of all, monetary amounts recovered by the agencies as a result of judgments or settlements are an explicit evidence of the work being done rather than the effect on the level of existing fraud in Medicare. It is unfortunate that cost of false negatives and the dynamic nature of fraud is not a major concern in the literature.

We outline two methods to detect possible fraud. We use clustering methodology to group zip code areas in terms of socioeconomic factors. We identify providers with outlying rates of utilization within these relatively homogenous regions. The second method includes a definition of an impractical distance traveled for health care services and we flag providers based on the measures involving the claims above this distance.

It is clear that many providers who are flagged in these studies are innocent of fraudulent activity and a secondary analysis that requires the participation of beneficiaries and providers is necessary to minimize overall costs to the system.

3. Data

The database consists of claims belonging to beneficiaries requiring the utilization of infusion therapy drugs. This database has several characteristics that have implications for any type of fraud investigation. First, there are three types of physician identifiers. Not all identifiers are uniquely associated with a provider. Furthermore providers can use a combination of these identifiers within their claims. In other words the identifiers are not consistently populated for all claims of an individual provider. Some group providers also have identifiers listed in these fields. For some of these fields, individual providers are not distinguishable. In addition carriers can have their own rules to assign one of these

identifiers. We develop a methodology to uniquely identify, as much as possible, the individual providers. This is not a trivial issue for analysts building fraud models.

Among the large number of available variables in the data, the ones we use for the methodologies exposed in this paper are provider and beneficiary zip codes, provider paid amount, provider and beneficiary identification keys and provider physical address. We obtain socio-demographic information in this paper from Census 2000 web site (http://dataferrett.census.gov).

Due to the sensitive nature of the data, we can not provide much of the detail. The database is unsupervised, in other words we do not have a historical account involving fraudulent claims and criminal providers. Note that the census bureau aggregates data per Zip Code Tabulation Areas (ZCTA), whereas the spatial information that we have relates to zip code areas. This problem has been noted by Grubesic and Matisziw (2006). We believe that this distortion between Zip Code and ZCTA does not invalidate our study since our goal is to identify outliers and not a precise measurement of distances.

4. Models

This section details the two models we use in order to flag providers. The first model makes use of clustering procedures as well as regression for geographical analysis of possible fraud. We create demographically homogenous zip code regions using clustering procedures. We associate each zip code region with a random variable that can discriminate between health care utilization or billing areas. We run regression analysis to achieve this discrimination and create homogenous healthcare utilization and billing cluster regions. For each homogenous region, we detect possible outliers in terms of rates of utilization or billing. We start building the second model based on distances that beneficiaries travel in a given day from the centroid of their zip code to the centroid of the provider's zip code. We define an impractical distance traveled based on this data. The flagging procedures are based on claims that have respective distances that are at or greater than this distance.

4.1. Peer comparison

Separate analysis for regions under a purview is not a new approach. For instance OIG report "Aberrant Billing in South Florida for Beneficiaries with HIV/AIDS" (2007) separately analyzed three counties in South Florida which accounted for 72% of submitted charges for services relating to beneficiaries with HIV although only 8% of the beneficiaries with HIV resided in these counties. In fact in the same report OIG recommends physical site visits specifically for these high risk areas.

The first section in the Peer Comparison model involves finding homogenous five digit ZCTAs in the region under our purview. We use the data from Census 2000, where there are a large number of candidate variables to choose from. We leave the exposition of this exploratory data analysis to another publication. Table 1 lists the variables used in this study.

We outline the first section of the model in five steps:

Step 1. Use clustering procedures to group areas of similar socio-demographic areas, designated by ZCTA labels.
Step 2. Transform these groups into dummy variables.
Step 3. Identify a quantity of interest which would help identify fraud. Use this quantity as an independent variable in order to regress to the dummy variables from step 2. Group together the clusters that were formed in step 1 but are not statistically significant into one large cluster.

Table 1
Variables chosen for clustering analysis.

Population
of households
Average house value
Income per household
Latitude
Longitude
Minority population
of people in age cohorts 0–9, 10–19, 20–44, 45–65, >65
Male/female ratio
Number of people below the poverty line
Amount of land area

Step 4. Repeat steps 2 and 3 with another random variable of interest.
Step 5. Associate provider's physical office ZCTA with the cluster. Sort the quantity of interest from highest to lowest and flag those providers who have more than 1.5* the interquartile range + the 75th percentile.

Khattree and Dayanand (1999) note that, there are multitudes of clustering algorithms with different criteria to quantify similarity between objects. Each algorithm can lead to a different number of clusters and objects within each cluster. The authors point out that before clustering begins, the analyst needs to take into consideration, the purpose of the study. This concept drives our choice in the selection of clustering algorithms. In effect, we choose the algorithms in order to create a reasonable number of clusters with approximately equal sizes.

de Graeff et al. (2001), Kuper et al. (2008) and Sun et al. (2009) provide ample evidence that socio-demographic factors are important drivers of health care needs. Furthermore we believe that geographical proximity of ZCTA regions should factor in as well. These are the primary reasons for the variable selection for the clustering procedures.

In proceeding with step 1 we state our first goal as constructing relatively homogenous regions in terms of their socio-demographic variables and location. Fig. 1 contains a simplified

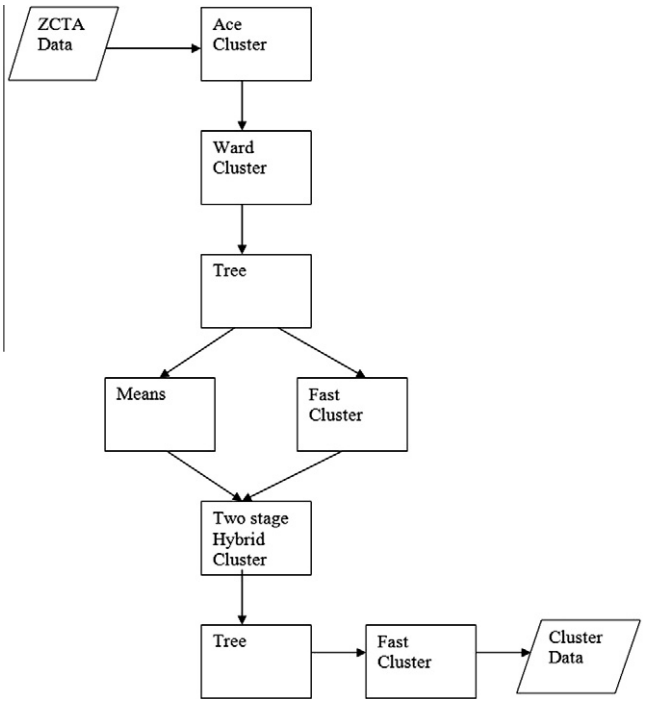

Fig. 1. Flow methods that create the initial clusters.

flowchart of the clustering and other necessary procedures for the final assignment of ZCTAs to their respective clusters. We choose these procedures and their sequence out of many other possibilities to satisfy these requirements. We obtain the values for the variables in Table 1 for each ZCTA in the region under our purview from the Census Bureau database. We remove from analysis those ZCTAs that have no recorded population as well as those who have either no income per household and no recorded house value. The steps in Fig. 1 are executed in SAS version 9.1.3. We use the Ace Clustering algorithm to estimate pooled within covariance matrix where knowledge of the number of clusters is not necessary. We retain the resulting canonical variables from the procedure in the Ward's clustering approach. The tree procedure in SAS allows us to retain the cluster membership from the Ward's clustering algorithm where we fix the number of clusters at 20% of the total number ZCTA regions. For each cluster we obtain the means of the canonical variables and use them as seeds in the fast clustering algorithm where the maximum number of clusters is set once again at 20% of the total number of ZCTA areas. The results indicating cluster membership from the fast cluster and the means of the canonical variables are merged to be used in the two stage hybrid clustering algorithm. Once again we set the number of clusters to 20% of the total ZCTA regions in the tree procedure to identify cluster membership. Fast clustering algorithm is used to identify the final cluster membership in three steps. In the first step we identify all clusters that have more than 10 ZCTA regions. The mean of these larger clusters become seeds for the second fast cluster analysis. The third and final application of the fast cluster assigns the remaining ZCTA regions to the clusters formed in the second step.

In order to associate these cluster groups with the quantity of interest we create dummy variables to denote cluster membership of ZCTAs and regress them against an independent variable. The goal of the regression analysis is to identify statistically significant ZCTA regions in terms of the independent variable. Once these similar ZCTA regions are constructed we can identify outliers of any random variable of interest. This will allow us to make the claim that in areas that are similar in terms of the independent variable, there are health care providers who are outliers of a quantity of interest. The clusters that are not statistically significant are merged with the reference cluster. For each statistically significant cluster, and the reference cluster group, we rank order the quantity of interest and flag a provider as an outlier whose quantity of interest in his/her respective cluster is above the median plus 1.5 times the interquartile distance. We show this set of rules in Formulation 1 where Q_i represents the quantity of interest of provider i.

Formulation 1.

Flag = 1 if $Q_i > C(P^{75}) + 1.5^*(C(P^{75}) - C(P^{25}))$
0 otherwise
Where $C(P^n)$ is the nth percentile of Q in cluster C.

4.2. Distance analysis

The second model is implicitly based on a subjective utility model. We assume that the utility of a beneficiary is composed of two attributes, the distance between the beneficiary and provider address, and the expected quality of service. We assume that utility of the beneficiary decreases with the increase in mileage between beneficiary and provider address. Unfortunately it is not possible to observe the expected quality of a provider from the perspective of a beneficiary. We hold that the disutility associated with distance traveled should be comparable for each individual. If this argument is accepted it leads to the point that in general, beneficiaries who travel relatively long distances do it for the high quality of provider services. Several other explanations are possible to account for relatively long distances. For instance a beneficiary address that is not updated in the data, a beneficiary who requires health services during vacation are common explanations that do not involve fraud. The model we construct accounts for many of these cases.

On the other hand, there are several types of fraud that we can hope to identify by investigating providers involving relatively long distances. For instance identification theft by unscrupulous staff members, billing agents, health care providers are such common agents of fraud. These agents through outright bribery or promise of free services to beneficiaries charge Medicare for services that are not medically necessary. In many cases the promised free services cause the beneficiaries to travel to locations farther than they normally would for such services. In other cases the beneficiaries are brought to the offices of the agents. Most such schemes of fraud are not sophisticated enough to change personal information of the beneficiaries or the providers to accommodate for the abnormal distances traveled to seek healthcare. The DOJ issued press releases involving the indictments against individuals who have committed this type of fraud "Three Arraigned in Medicare Fraud Scheme (2005) and Los Angeles Area Health Care Company Owner Pleads Guilty to Medical Identity Theft and Medicare Fraud" (2008).

Beneficiaries in rural and urban areas do not have the same access to health care. In general, beneficiaries in rural areas have to travel longer distances on average than the beneficiaries in urban areas for health care needs. As we have a relatively small amount of data from beneficiaries living in rural ZCTA regions, we decided to concentrate on beneficiaries living in Metropolitan Statistical Regions (MSA). MSA regions can include rural regions but only if they have strong social and economic ties to urban regions. These ties are a function of employment and commuting between these rural and the urban areas. If a beneficiary is from a non-MSA region we remove his/her claims from the database. We construct four distance brackets where we define these as the median – 75th, 75th–95th percentile, 95th–99th and greater than the 99th percentile of distances traveled. To prevent cases of beneficiaries who reside out of state skewing the distances for these percentiles, we compute the distances traveled between beneficiaries and providers only within the state under purview. For every healthcare provider who earns above $50,000 we compute the proportion that the provider earns from beneficiaries who traveled distances in between these thresholds that we refer to as, threshold brackets. For each distance threshold bracket, we take the 95th percentile of the providers' above mentioned proportions. We refer to these four numbers as the earnings thresholds. Furthermore we define abnormal distance, as the distance greater than the 99th percentile of distance travelled between the beneficiary and the provider address. In order not to skew these percentiles we do not count multiple claims by the same provider for a beneficiary on a single day. These two sets of statistics allow us to construct a flagging procedure. We flag a provider if he/she has more than 10% of his/her beneficiaries who traveled above the 99% threshold or he/she has the respective percentage of earnings more than the two earnings thresholds. We show this set of rules within Formulation 2.

Formulation 2.

Flag = 1 if sum $(I(Q_{i,(50-75)} > \pi_{(50-75)}), I(Q_{i,(75-95)} > \pi_{(75-95)}), I(Q_{i,(95-99)} > \pi_{99}), I(Q_{i,(99-100)} > \pi_{(99-100)})) > 2$
or
$B^{99} > .1^*B$
0 otherwise

where I is an index function which evaluates to 1 if true 0 otherwise,

π_{n-m} is the 95th percentile of proportion of earnings obtained from services to beneficiaries whose traveled distance is within the nth to mth percentile distance traveled.

B^{99} represents total number of beneficiaries of a provider who travelled more than the 99th percentile of beneficiaries travelled distance.

In order not to multiple count the number of times the beneficiary travels to the provider location we concatenate the beneficiary identifier with the date of service in order to distinctly select this variable in computing the distance between the two zip codes. The distance computation between the centroids of Zip Codes is the well known Haversine Formulation.

5. Demonstration

This section illustrates the two methods that Section 4 describes by providing hypothetical examples for each method. Sections 5.1 and 5.2 provides hypothetical examples for the peer comparison model of 4.1 and 4.2, respectively.

5.1. Demonstration peer comparison

Assume that the region under purview is composed of Medicare beneficiaries in state *X*, *Y* and *Z*. The question that naturally arises is whether the clustering procedures should be done for the states separately or together. We choose to pool the ZCTAs together as it is our contention that the information provided by state label does not provide useful information for the purpose of clustering socio-demographic data when the variables in Table 1 is known. The clustering procedures in Section 4.1 lead to *N* cluster being formed.

The next step is to merge the quantity of interest, which for the purposes of the demonstration is dollar amount paid per beneficiary paid to the provider, with the ZCTAs. The amount paid per beneficiary becomes the dependent variable of our regression analysis. The *N* number of clusters becomes the independent variables represented with $N-1$ dummy variables. Table 2 shows the data before regression analysis.

The regression analysis merges the statistically non-significant clusters into the reference cluster using a backward regression model. A statistically significant cluster implies that the membership to that cluster implies a degree of knowledge on the variable, amount per beneficiary paid to the provider.

What we obtain as a result of the procedures outlined above is a number of clusters of ZCTA regions which are relatively homogenous socio-demographically and statistically significant in a regression model where the dependent variable is the amount claimed per beneficiary by the provider. In other words, these clusters, in addition to being relatively socio-demographically homogenous, they are also relatively homogenous in terms of amount paid per beneficiary for a provider. The reference cluster that in many similar applications will not be statistically significant and is composed of clusters that is found not to be statistically significant during the iterative backward elimination process constitute a large number of ZCTAs. This means that the reference cluster can not be ignored. Using Formulation 1 we flag outlier providers for all statistically significant clusters and the reference cluster.

5.2. Demonstration distance analysis

After the elimination of ZCTAs that are not associated with Metropolitan Statistical Areas (MSA), the next step is to compute the daily distances traveled between the provider and the beneficiary. Each claim that has been submitted contains five digit Zip Code locations. Table 4 shows three rows from a hypothetical database where a provider has three total claims. In this table the provider has two of these claims provided for the same beneficiary on the same day. The claim types are different and hence the two claims. Therefore we assume that beneficiary XXXX has made a single trip on this day and reflect this on the computations involved in the distance analysis. If Table 3 contains all the claims of provider 000000001, his/her average distances traveled is 2.5 miles.

Table 4 shows the result from the tabulation of the claims in Table 4 and reports the distances traveled on each day between a beneficiary and provider address.

Once we have all the claims of all the providers we go onto compute the percentiles of distances traveled. Assume that these are 2, 5, 25 and 49 miles for 50th, 75th, 95th and 99th percentiles,

Table 2
Claims data with cluster membership and canonical variables.

Amt/bene.	Cl. 1	Cl. 2	...	Cluster $N-1$	Canon. 1	...	Canon. 14	Zip code
$100	1	0	0	0	.14324	...	.511	99997
$125	1	0	0	0	.43533	...	.522	99997
$222	0	1	0	0	−.324	...	.234	99992
...	.	.	.	.	...	...	...	...

Table 3
Distance between the centroid of beneficiary and provider zip codes, concatenated service date and beneficiary ID to distinguish distinct beneficiary trips.

Bene. zip C.	Prov. zip C.	Distance	Date of service‖Bene. ID	Prov. ID	Bene. ID	Claim type
99997	99992	2.5 miles	04082008XXXX	000000001	XXXX	1
99997	99992	2.5 miles	04082008XXXX	000000001	XXXX	2
99997	99998	2 miles	04072008XXXY	000000001	XXXY	2

Table 4
Tabulated results with distinct beneficiary trips.

Bene. zip C.	Prov. zip C.	Distance	Date of service‖Bene. ID	Prov. ID	Bene. ID
99997	99992	2.5 miles	04082008XXXX	000000001	XXXX
99997	99998	2 miles	04072008XXXY	000000001	XXXY

Table 5
Proportion of earnings within each distance bracket.

Provider ID	2–5 miles	5–25 miles	25–49 miles	>49 miles
000000001	.5	.1	.05	.1
000000002	.3	.05	.1	.03
000000003	.6	.1	.2	.04
⋮	⋮	⋮	⋮	⋮
000000010	.4	.1	.01	.02

respectively. Table 5 shows a hypothetical case involving the proportion of earnings for each *distance bracket*. The sum of the row values in Table 5 will be equal to one minus the proportion of earnings the provider has from the beneficiaries who travel less than the median amount of distances traveled, computed from a database in the form of Table 4.

We proceed to obtain the 95th percentile of the columns in Table 5 and determine earnings threshold values. If a provider has any more than two of his/her proportion of earnings within the distance brackets listed in Table 5 then he/she is flagged. Independent of this rule we also flag the provider if more than 10% of his/her beneficiaries arrive from distances that are greater than the 99th percentile (49 miles) of distances traveled.

6. Results and conclusion

In this paper we listed two methods to identify fraud in Medicare infusion therapy providers. We do not make the claim that every health care provider flagged in this study is involved in fraud. On the other hand, we argue that these methodologies which identify possible cases of fraud are useful as the first line of defense mechanism in identifying fraud. We believe a system dynamic approach is required to investigate the Medicare system in decisions involving the investigation of possible providers of fraud. We suggest that a provider who gets flagged as a result of an unsupervised fraud detection method needs to go through a second stage that should cause him/her the least disutility.

References

Anderson, G. F., Frogner, B. K., & Reinhardt, U. E. (2007). Health spending in OECD countries in 2004: An update. *Health Affairs, 26*(5), 1481–1489.

Anderson, G., & Hussey, P. S. (2001). Comparing health system performance in OECD countries. *Health Affairs, 20*(3), 219–232.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–249.

de Graeff, A., de Leeuw, J. R., Ros, W. J., Hordijk, G. J., Blijham, G. H., & Winnubst, J. A. (2001). Sociodemographic factors and quality of life as prognostic indicators in head and neck cancer. *European Journal of Cancer, 37*, 332–333.

Department of Health and Human Services Office of Inspector General Aberrant Billing In South Florida For Beneficiaries with HIV/AIDS. <http://www.oig.hhs.gov/oei/reports/oei-09-07-00030.pdf>. Accessed 19.03.09.

The Department of Health and Human Services And The Department of Justice Health Care Fraud and Abuse Control Program Annual Report For FY 2007. <http://www.oig.hhs.gov/publications/docs/hcfac/hcfacreport2007.pdf>. Accessed 18.03.09.

Detailed Information on the Health Care Fraud and Abuse Control Assessment. <http://georgewbush-whitehouse.archives.gov/omb/expectmore/detail/10000292.2002.html>. Accessed 19.03.09.

Grubesic, T. H., & Matisziw, T. C. (2006). On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *International Journal of Health Geographics, 5*(58), 1–15.

Hoffman, E. D., Klees, B. S., & Curtis, C. A. (2008). Brief summaries of Medicare & medicaid title XVIII and title XIX of the social security act as of November 1, 2008. Center for Medicare and Medicaid Services. <www.cms.hhs.gov/MedicareProgramRatesStats/Downloads/MedicareMedicaidSummaries2008.pdf>. Accessed 13.03.09.

Khattree, R., & Dayanand, N. N. (1999). *Applied multivariate statistics with SAS software* (2nd ed.). NC: SAS Institute Inc..

Kuper, H., Polack, S., Eusebio, C., Mathenge, W., Wadud, Z., & Foster, A. (2008). A case-control study to assess the relationship between poverty and visual impairment from cataract in Kenya, the Philippines, and Bangladesh. *PLOS Medicine, 5*(12), 1716–1728.

Li, J., Huang, K. Y., Jin, J., & Shi, J. (2008). A survey of statistical methods for health care fraud detection. *Health Care Management Science, 11*(3), 275–287.

Los Angeles Area Health Care Company Owner Pleads Guilty to Medical Identity Theft and Medicare Fraud. <http://www.usdoj.gov/opa/pr/2008/October/08-crm-927.html>. Accessed 26.03.09.

Performance and Management Assessments Budget of the United States Government Fiscal Year, 2004. <http://www.gpoaccess.gov/usbudget/fy04/pdf/pma.pdf>. Accessed 15.03.09.

Sun, X., Rehnberg, C., & Meng, Q. (2009). How are individual-level social capital and poverty associated with health equity. A study from two Chinese cities. *International Journal for Equity in Health, 8*(2), 9–22.

Three Arraigned In Medicare Fraud Scheme. <http://www.usdoj.gov/usao/cac/pressroom/pr2005/113.html>. Accessed 26.03.09.

United State Attorney's Office. District of Rhode Island. <http://www.usdoj.gov/usao/ri/press_release/april2008/wehbe_forfeiture.html>. Accessed 12.04.09.