MOX-Report No. 49/2018

Data Mining Application to Healthcare Fraud Detection: A Two-Step Unsupervised Clustering Model for Outlier Detection with Administrative Databases

Massi, M.C.; Ieva, F.; Lettieri, E.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Noname manuscript No.
(will be inserted by the editor)

Data Mining Application to Healthcare Fraud Detection: A Two-Step Unsupervised Clustering Model for Outlier Detection with Administrative Databases

Michela C. Massi · Francesca Ieva · Emanuele Lettieri

Received: September 2018

Abstract Being the recipient for huge public and private investments, the healthcare sector results to be an interesting target for fraudsters. Nowadays, the availability of a great amount of data makes it possible to tackle this issue with the adoption of data mining techniques. This approach can provide more efficient control of processes in terms of costs and time compared to manual audits. This research has the objective of developing a novel data mining model devoted to fraud detection among hospitals. In particular, it is focused on the DRG upcoding practice, i.e. the tendency of coding within Hospital Discharge Charts (HDC) in Administrative Databases, codes for provided services and inpatients health status so to make the hospitalization fall within a more remunerative DRG class. The model here proposed is constituted by two steps: one first step entails the clustering of providers according to their characteristics and behavior in the treatment of a specific disease, in order to spot outliers within this groups of peers; in the second step, a cross-validation is performed. This second phase is useful for controllers to verify whether within the list of suspects identified in the first step, any hospital exists, which may be justified in its outlierness by its particular characteristics, or by the treatment of a more complex patients' base. The proposed model was tested on a database relative

Michela C. Massi
Department of Management, Economics and Industrial Engineering
Politecnico di Milano
Lambruschini Street 4/c, 20100 Milano, Italy
Tel.: +39-02-23994077
E-mail: michela.massi@mail.polimi.it

Francesca Ieva
MOX - Laboratory for Modeling and Scientific Computing
Department of Mathematics
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
Tel.: +39-02-23994578
E-mail: francesca.ieva@polimi.it

Emanuele Lettieri
Department of Management, Economics and Industrial Engineering
Politecnico di Milano
Lambruschini Street 4/c, 20100 Milano, Italy
Tel.: +39-02-23994077
E-mail: emanuele.lettieri@polimi.it

to HDC collected by Regione Lombardia (Italy) in a time period of three years (2013-2015), focusing on the treatment of heart failure.

Mathematics Subject Classification (2000) MSC 62P10 · MSC 90B50

Keywords Data Mining, Hospital, Fraud, DRG, Upcoding, Administrative Database

1 Introduction

Healthcare is, by design, subject to frauds [1]. These frauds have different perpetrators (e.g. hospitals, medical figures, private facilities) and different dynamics. The three most relevant types of fraud are upcoding, cream skimming, and unbundling [2].
The *upcoding* practice consists in classifying a patient in a DRG, or registering treatment codes (in case of Fee for Service[1] payment systems) that produce higher reimbursements [4]. This practice by public hospitals is more likely to be due to unintentional errors by coders, or misunderstandings with doctors; while, when talking about private hospitals or medical practices, it could actually be due to profit maximising purposes [5, 6]. *Cream skimming* consists in some discrimination
among patients, in order to select the most lucrative ones. Providers may indeed achieve higher productive efficiency by focusing on easy to treat patients, treating more patients with less severe conditions and attaining higher profits [2, 7–9].Finally, the term *unbundling* refers to the practice of using two or more DRG codes instead of one more inclusive code. It also refers to the practice of billing multiple times in order to obtain higher reimbursements for procedures that could have been submitted as a single bill [10].
Given that Healthcare is the target of large public and private investments (on average, in OECD countries, 15% of the governament budget is allocated for this purpose), independently from the geographical and political setting, this sector is a rather appetible one for frauds. A 2017 OECD report lists some of the major worldwide frauds: EUR 200M in France, GB 11.9M in the United Kingdom, USD 2.4B in the United States [11]. In this ever-growing healthcare industry, using manual countermeasures to fight frauds is not enough. Taking as an example the Italian setting, the operations currently performed by the controlling mechanism (Nucleo Operativo di Controllo, NOC) in Lombardy Region, which is one of the most populated regional districts in Italy and the leading one in terms of Healthcare expeditures, entail the manual scrutiny of 14% of the yearly registered Hospital Discharge Charts (HDC, or SDO in italian - "Scheda di Dimissione Ospedaliera"), mostly selected randomly (Regione Lombardia, 2018). The only attempt to restrict the scope of this investigation regards the monitoring of certain parameters, intervening in those cases where acceptable thresholds are surpassed, registering hospitalizations with evidently suspicious incidences (La Repubblica, 2008).
Given the ever-growing availability of digital data, the adoption of data mining techniques might help to reach better results and more efficient processes, in terms of both time and costs. Developing tailored algorithms would allow for the restriction of the pool of investigated providers, including those acting cautiously and perpetrating fraud within the limits of the monitored indicators [12].
In this paper, we deal with the development of a novel systematic and quantitative approach fraud detection, with a focus on the upcoding practice, because of the extremely relevant economic impact this kind of fraud has on the system [13]. The objective is proposing a novel tool

[1] A method in which doctors and other health care providers are paid for each service performed. Examples of services include tests and office visits. [3]

to support human decisions in the preliminary phases of screening providers to spot suspects eligible for a more in depth investigation, including those with a more cautious approach to fraud. Because of the large availability of Administrative Databases, we decided to exploit this type of data as our source of information. All the analysis described in the following Sections were performed using R [14].
The proof of concept of our method, which is generally adaptive to any kind of system, is was developed in the Italian context, the Lombardy Region in particular, studying the behavior of hospitals adopting regional administrative data. As mentioned before, this region represents a relevant benchmark in the Italian healthcare landscape and as such it represented a promising starting point for our testing, but the flexibility of the developed method allows for transferring it to other contexts.
The article is organized as follows: in Section 2 we analyzed the existing studies in terms of healthcare fraud, and the researchers' attempt to tackle it through data mining techniques; Section 3 will then extensively go through the whole Model Design process, from data extraction criteria (focused on Heart Failure, to solidify the statistical reliability of this model), to the pre-processing and data preparation phase, to the conceptualization of the model itself (Subsection 3.1). In Section 4 we test the designed methodology on a dataset provided by Lombardy Region, and we provide some numerical insights. Finally, in Section 5 we discuss such results, and conclude with some final remarks on the value added of this research, and the impacts it might have on policy making.

2 Literature Review

Literature on the subject of healthcare fraud is mainly focused on discussing empirically causes and consequences of healthcare systems and fraudulent behavior (e.g. [15–18]), more than searching for a way to detect such fraud. Data mining approaches to the problem, even though they are gaining momentum, are still at their infancy.
Even though still limited, some contributions to the literature in the field of data mining applied to healthcare fraud detection exist, most of them coming in the form of conference papers. We can group the existing methodologies of fraud detection into three main groups: supervised, unsupervised, or hybrids of the two [19]. Supervised techniques [20–23] despite their undeniable potential and predictive power, exhibit the risk of focusing on old patterns and losing predictive capability as new records are evaluated over time [28]. For this reason, unsupervised techniques are the most adopted [12, 24–27, 29–35, 43]. This indeed is a reasonable approach overcoming the aforementioned limitation of supervised methods, since the intrinsic characteristic of fraud is that of changing in time, according to the arising regulations or control systems. Hybrid techniques or on-line processing systems take the best of both approaches [19, 36, 37], creating a combination of supervised and unsupervised methods. This is surely an interesting field of study which still deserves lots of attention by future researchers. However, for the sake of this research, the focus remains the one of unsupervised techniques, because of their simpler development and applicability.
As previously mentioned, this paper focuses on the upcoding practice, that can be declined in several ways and has drawn the attention of most data mining researches because of the extremely high impact it has on healthcare expenditures. The majority of the available literature attempts to spot providers with very high claiming episodes, which distinguish themselves as *evident* outliers - one clear example in [35]. However, the evolving nature of modern fraud, has pushed some researchers to change direction and try and identify providers that have a more cautious approach to fraud, which would be neglected in the aforementioned high-claiming

groups. The *behavioral* models suggested by Musal *et al.* [12] and Shin *et al.* [31] try indeed to respond to this objective. However, those techniques have still very limited application, even though their usefulness in reducing the overall level of fraud in the system is undeniable. The lack of interest from most researchers can be justified by the intrinsic risk of these models to exhibit higher false positive rates, and the lower amount of recovery from each correctly spotted fraudster [12]. However, as reported in [31], types of fraud are growing increasingly sophisticated, and patterns detected from fraudulent and non-fraudulent behaviors become rapidly obsolete because of rapid changes in behavior, and fraudulent providers are becoming smarter in finding more cautious approaches which usually prevent them from being investigated [38]. In conclusion, from the study of previous contributions emerges the need for more data mining methods capable to adapt to changing behaviors of fraudsters, capable of identifying more cautious approaches to fraud. For this reason, the model proposed in this study belongs to this more complex and less tackled pool of unsupervised methodologies, aiming at studying providers with a *behavioral* approach, to recognize the most subtle fraudsters' patterns.

3 Methods

The goal of this research is developing a tool capable of supporting the preliminary phases of healthcare fraud's investigation for upcoding episodes, with a particular attention to cautious behaviors. In order to do that, our approach needs to properly preprocess the available information. Adopting Administrative Databases, a couple of remarks need to be highlighted about how upcoding is perpetrated and recognizable within such datasets. When a patient is hospitalized by a provider, such provider takes charge of its care, performs a series of diagnosis and prescribes a series of treatments or surgical operations, which are registered in the HDC, and altogether affect the final Diagnosis Related Group (DRG) of the single hospitalization once elaborated by the *grouper*. The concept of *upcoding* is strictly related to this phase of the care process. Hospitals and healthcare providers in general, may have the incentive to modify the registration of the diagnosis or treatments the patient was provided with, in order to make the hospitalization fall into a more expensive DRG category. For example, coding complications (CC) in the Principal Diagnosis, will drive the value of the reimbursement for such HDC higher.

DRGs can therefore be considered a good proxy of hospitals' behavior. Since every HDC contains information regarding both diagnosis, comorbidities, treatments, and patients' characteristics, it can be stated that DRGs are an aggregate reflection of each case the provider managed.

Indeed, it can be assumed that similar hospitals (in terms of e.g. dimension, ownership, and specialization), should face similar casemixes, and behave (therefore coding in the HDCs submitted for reimbursements) similarly: they should face an analogous set of cases, thus coding for a comparable set of diagnoses, treatments and patients' severity.

Therefore, in order to design a model capable of detecting suspicious yet cautious providers, there was the need to obtain from a source of HDC data, complete information about (i) the behavior of the hospital, (ii) the characteristics of such providers and (iii) the clinical history of the patients treated in the different facilities. Behaviors and characteristics of providers are needed to spot suspects among similar peers, while the pool of the treated patients and the estimated complexity of their care can be adopted to validate the identified outliers, verifying whether the highlighted discrepancies can be due to a higher incidence of complex cases faced by the provider. This should help in reducing the risk of false positives in the process of supporting the preliminary scrutiny.

This being the rationale behind the developed methodology, the first step to obtain the needed information was an extraction phase. Our model was developed and tested on Administrative Data (originated from HDC) about all hospitalizations for heart failure happened in the Lombardy Region within the timespan 2013-2015. Patients' data were already anonymized and both patients' and hospitals' confidentiality were preserved,since their knowledge was not relevant for this research. In fact, this study aimed at developing a novel methodology, and not to spot any particular actor, or to substitute the traditional auditing activities.
We aimed at developing a transferrable tool, applicable to all diseases in the healthcare domain, but to improve its performance and proving the concept, the initial scope was reduced to the treatment of a single disease. The selected disease was Heart Failure (HF), and the initial extraction phase (as reported in the pseudo code in Algorithm 1) was performed accordingly. First, all HDCs with a subset of Heart Failure-related ICD-9 codes in any position (primary diagnosis, secondary diagnosis, etc.) have been extracted (the complete list of the adopted codes is available in Appendix). At this point, for all patient included in the extraction, his/her HDCs for any kind of treatment or disease was extracted.
As mentioned before, the choices behind this extraction phase had the objective of allowing

Algorithm 1: Initial HDCs extraction

```
Input  : Whole HDC Dataset
if HF-code in DIAGX_ID then
    Heart_Failure = 1;
else
    delete;
end
Output: HF Dataset
Input  : Whole HDC Dataset
if ANONYM_ID in HF Dataset then
    keep;
else
    delete;
end
Output: Final HDC Dataset
```

us to gather a complete view of both hospitals and patients. However, in order to feed the model, such data had to be reshaped and aggregated into different datasets, all with a different granularity and statistical unit, to obtain the three sets of complete information highlighted earlier.

- The *HDC Dataset*, resulted directly from the extraction. This dataset has one row per hospitalization (reported in the form of a full HDC with information about the patient and the treatments or diagnosis he/she encountered). Each HDC was subsequently associated with its Comorbidity Index, as an estimation of its complexity. To calculate the Comorbidity Index for each HDC in the first place, the algorithm for the Combined Comorbidity Score by Gagne *et al.* (2011) was adopted. This method provides a single numerical score associated to an hospitalization episode, using informations regarding the coded comorbidities within the HDC [39].
- The *Patients Dataset* was built aggregating all available information within the HDC Dataset at patients' level. Each row represents one of the patients that experienced an HF episode in the time span of the available data. For each patient, the available information within all their HDC was aggregated, reporting variables such as age at their first hospitalization,

total number of hospitalizations, sex, average lenght of stay, etc. One variable of particular interest is the Comorbidity Index, estimated cumulating for each patient the comorbidity scores obtained as mentioned before for each HDC, in order to estimate the health status of the subject, and the derived complexity of his/her care [39].
– The *Hospitals Dataset* reported all information with the single hospital as statistical unit (one hospital per row). This dataset required a more extensive procedure, with the aim of including all relevant information to identify a fraudulent behavior by the providers. Using the variables found within literature as a reference, all information directly available within the HDC dataset useful to identify fraud was aggregated for each hospital (e.g. cost per hospitalization [2, 12, 25, 26, 35], average length of stay [9], number of episodes in a given period of time [35], etc.). Then, a set of additional indexes were estimated, as detailed in the following.

The first estimated index is the *Upcoding Index* as expressed in the study conducted by Silverman and Skinner [5]. The authors considered a group of four DRGs related to *pneumonia* and calculated the *Upcoding Index* as the ratio between the most remunerative DRG (DRG 79), and the sum of all the four DRGs related to this disease. In order to repurpose this index in our context, several assumptions were made. The Tariffs' List provided by the Healthcare Ministry was used as reference, and only heart failure-related DRGs were selected. The Pareto Curve of DRGs' reimbursement cost was traced, placing the set of selected DRGs in descending order and selecting those DRGs that together accounted for the 60% of the value of all HF DRGs all together. The selected DRGs were 7, out of 44.
Once selected the *heaviest* DRGs in terms of reimbursements received by healthcare providers ($heavyDRG_i$), the rate of incidence of those DRGs on the overall registrations for HF cases was computed (DRG_i) for each hospital ($i = 1,...,I$) as:

$$UpcodingIndex_i = \frac{\sum heavyDRG_i}{\sum DRG_i} \tag{1}$$

The second indicator we decided to include is another Upcoding Index, in the version proposed by Berta *et al.* [2]. According to the authors, the index can be estimated as in (2). This version of the index was inspired by Silverman's work, improved with the additional 'Comorbidity' load, which adjusts results taking into consideration patients' illness status.

$$UPCODING_{it} = \frac{S^C_{it}}{S^C_t} * \frac{1}{CI_{it}} \tag{2}$$

Being S^C_{it} the share of discharges with complications over the total number of discharges in hospital i at time t. This share is compared to the share computed at regional level (the denominator factor). The ratio shows whether hospital i at time t is treating more complex cases than regional average [2]. The ratio is then divided by the Comorbidity Index.

In order to calculate the new *Upcoding Index* in (2) with the data at hand, it was decided to select all DRGs related to HF, since the authors mentioned the '*major disease category*' as a discriminant factor. Another assumption was needed, regarding the *Comorbidity Index*: the *average comorbidity weight* of treated patients was adopted as a proxy of that value for each hospital.
To select the HF-DRGs with complications (DRG with CC), the selection was based on the official documentation provided by the Health Ministry about DRGs and relative tariffs, selecting all DRGs where 'CC' was explicitly reported in description. In addition, DRGs with AMI (*Acute Myocardial Infection*) were included in the selection as well, since the presence of this disease is a discriminatory factor amongst DRGs with similar descriptions.

A particular attention was then devoted to the work of Ekin *et al.* [40], since their contribution suggested a way to represent providers' behavior, instead of estimating a single measure - accordingly to the objectives of this research.
Their method is based on Concentration Functions, an extension of *Lorenz curves* (see e.g. [41]), used to compare providers among each other (and in respect to the whole group) in terms of probability of occurrence of a set of common base of specific treatments. In other words, the authors select a common base of billed treatments of a group of providers, and they calculate a set of probabilities for each of these providers and treatments, i.e. p, q and r, where

- q_i represents the probability of the whole population to bill for the treatment i (in their paper, the treatments' bills are called *outcomes* of the statistical experiment);
- p_{ij} is the probability that the hospital j bills the treatment i, while
- r_{ij} is the ratio between p_{ij}, and q_i..

What authors are interested to observe is the distance of p, to the group distribution q. The concentration function is then constructed as the cumulative sum of the probabilities of the *outcome* or treatment i (q and p), ordered according to the descending value of r_{ij}. The process of readapting this model for our objectives started by selecting HF-related DRGs within our dataset, in order to restrict the number of cases to a unique field and environment, as suggested by the authors. However, what made the concentration functions comparable in the study in [40], was the common base of treatments. In our specific setting, this approach was impossible, since no common base of DRGs existed among all considered hospitals. Therefore, the decision was to build the model neglecting such common base: this made curves differ in terms of origin and ending points, and less comparable in respect to the original model, but allowed to compare hospitals on the basis of which DRGs they did and did not register in the considered period. As explained at the beginning of this section, DRGs coding behavior can represent a good proxy of providers attitude in treating the disease under consideration. Therefore, for each HF-related DRG (i) in each hospital (j), the r_{ij} value was estimated as

$$r_{ij} = \frac{\frac{\sum DRG_{ij}}{\sum DRG_j}}{\frac{\sum DRG_i}{\sum DRG}} i = 1, ..., I j = 1, ..., J \tag{3}$$

The string of r_{ij} values for each hospital represented how the hospital behaved in terms of coding treatments for HF-affected patients.

3.1 Model Proposal

Given the outcome of the data preparation phase, we are now able to propose a two-step model for screening hospital behaviour and detecting potential upcoding fraud.
The first step aims at studying hospitals' behavior and recognizing potential fraudsters, the second aiming at supporting auditing decisions by providing a dashboard of informations useful to verify whether the identified suspects might be eligible for an in depth investigation.

Step One. This step exploits the r measure in (3) to discriminate among hospitals on the basis of how they behaved in the treatment of HF-affected patients. Healthcare providers differ significantly among one another. This difference can be due to several factors, such as dimension, technological endowment, resources, specialization and characteristics of the patients' population within its catchment area. Therefore, it would be meaningless to compare all hospitals together, and expect that they behave similarly. This considerations lead to the exploitation of a clustering algorithm after the computation of the aforementioned probabilities and characteristics' indicators: this would help in grouping similar providers, and identifying potential fraudsters among group of comparable peers. For this reason, this step entails the application a *K-mean* algorithm [42]. The entire process followed in Step One is reported in the pseudo code of Algorithm 2. The number of groups are defined on the basis of the Within Cluster Sum of Squares Method. Then, hospitals are grouped on the basis of their characteristics, the fraud-relevant variables and their behavior in the treatment of HF (represented by the different values of r): this allows to distinguish homogeneous groups of providers. After accounting for structural differences, and the different cases they may face, hospitals behaving differently from the peers they should be aligned with can be spotted. In order to detect outlying providers, the distance between each hospital and the representative of its corresponding cluster is computed. The hospitals with a distance above a specified threshold (in this case 95^{th} percentile) are selected and flagged as outliers.

Algorithm 2: Step One

Input : Hospitals dataset
Within Clusters Sum of Squares method;
Optimal number of clusters estimation (n);
k-means algorithm;
Output: n clusters (C) with n centroids (c)
Input : i Hospitals and n clusters and centroids
foreach C;
foreach $i \in C$;
Compute $dist(i, c)$;
$Distance \leftarrow dist(i, c)$;
endfor;
endfor;
Output: $dist_{iC}(i, c)$, $Distance$
$t \leftarrow 95_{th}$ percentile of the distribution of $Distance$;
Input : $dist_{iC}(i, c)$, t
foreach $i \in$ Hospital Dataset;
if $dist_{iC}(i, c) > t$**then**
  $FlagOutlier = 1$;
  $Suspects[x] \leftarrow i$;
else
  out;
end
Output : $Suspects[x_1, ..., x_m]$

Step Two The second step takes place after the identification and listing of all outliers. This phase is useful for controllers to create a visual dashboard that supports an informed skimming of the suspects identified in the first phase. The subset of identified outliers that behave in a way different from their comparable peers has a higher probability to deserve an in-depth investigation by human controllers. However, because *behavioral methods* demonstrate to have higher false positive rates - even though stronger in identifying cautious fraudsters [12] - a further validation of results is recommended. In order to support the controllers in this further validation, the tools presented in this second step were developed. A first operation for this

Variable	Dataset	Notes
Age	Patients	Indicator of patients' complexity
length of Stay	Patients, SDO	Indicator of patients' complexity
Comorbidity	Patients, SDO	Indicator of patients' complexity
Total Costs	Patients, SDO	Patients' expensiveness
Cost / length of Stay	SDO	Expenses in relation with intensity of care
Cost / Comorbidity	Hospital	Expenses in relation with intensity of care

Table 1: Variables of interest for cross validation

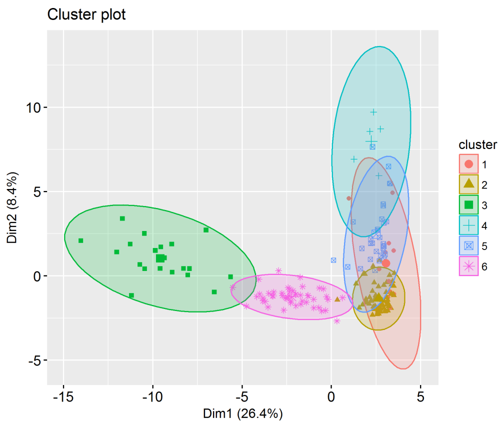


Fig. 1: Clusters resulting from applying *k-mean* algorithm to Hospitals dataset.

step entails using 4 variables describing hospital characteristics, which may be useful to justify their distance from their peers: degree of specialization, percentage of DRGs with complications (CC), Upcoding Index in (2) after the described adjustments, and average costs. Visualizing the values of those measures for each outlier, compared to the distribution among the peers in their cluster, would help collecting insights for informed decisions.
The second operation of this Step is a procedure of cross-validation with the patients' population treated by the hospital. To perform this validation, Patients Dataset came into play. The idea behind is to verify whether the anomalous behavior may be justified by the complexity of the treated population. To perform this passage the variables in Table 1 are used. Adopting once again visualization techniques together with t-tests, it is possible to provide decision makers with an easy and understandable tool to support their decisions. In this case the distribution of the variables is computed and plotted, allowing for the comparison of one specific hospital's behavior compared to the entire population. In the 'Dataset' column of Table 1 are listed all the data sources used to estimate those values. For instance, the Length of Stay (LOS) is computed both for each patient, and for each HDC (SDO) registered by the hospital.

4 Results

From the application of the clustering algorithm on the data at hand – after the estimation of the optimal number of clusters with the WSS method – the six clusters of hospitals represented in Figure 1 were discovered. The silhouette of these clusters is not particularly high (0.23, as

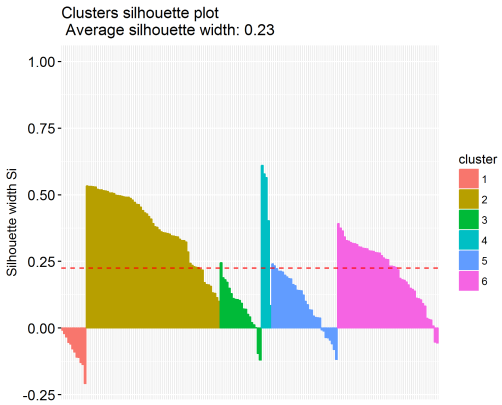


Fig. 2: Silhouette plot for the 6 clusters identified by the k-mean procedure.

shown in Figure 2), due to the high number of dimensions considered by the algorithm (64 variables). Anyway, it has been considered as satisfactory in view of the complexity of the problem and the results it allowed for. In facts, while evaluating clusters characteristics, clear differences among clusters were identified in some of the dimensions of interest.
Once the six clusters have been defined, the model requires the calculation of the local outliers for each cluster.In Table 2 the summary statistics of the local distances have been highlighted. The distribution function is represented in Figure 3. The function being so skewed demonstrates how most of the data points lie close to the center of their cluster, thus sustaining the goodness of the obtained clustering. Once again, this results goes in favor of our procedure, despite the low silhouette value. 10 hospitals resulted as outliers adopting the 95^{th} percentile threshold. Among these, 8 were private providers, and 2 public.

Variable	**Min**	**5%**	**1st.Qu.**	**Median**	**Mean**	**3rd.Qu.**	**95%**	**Max**
Local Distance	0.1913	0.2493	0.4264	0.6261	0.7303	1.0100	1.5127	1.9760

Table 2: Local Distances Distribution Summary

Three hospitals were selected for an exemplificative application of the second step (H31 - public -, H51 and H11 - private): all of them were the most distant outlier of their three different clusters, and they were interesting because of the different characteristics for example in terms of ownership, number of treated patients, average reimbursements etc. These three hospitals have been analyzed by evaluating the dimensions defined by the model. In Figure 4 all clusters and identified outliers are represented: each graph reports one of the four dimensions (described in Section 3.1) which characterize the hospital.

As for the cross-validation with the patients' population, each of the dimensions listed in Table 1 was evaluated, while just some of the most interesting results are reported here. Figure 5 reports one of the analysis about the length of the stay (LOS) -considered a proxy of complexity of the patient: the more the distribution is skewed to the right, the more the outlier faced a set of complex cases. Table 3 shows the values of comorbidity (another indicator of a

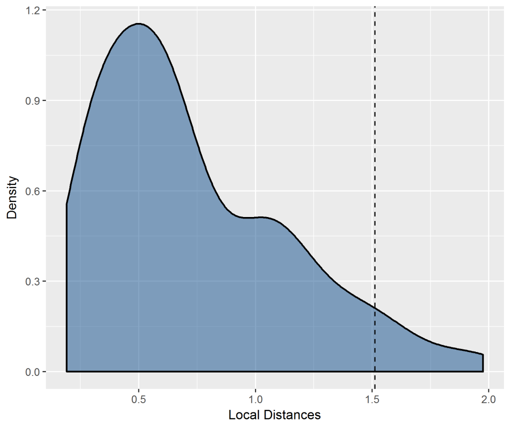


Fig. 3: Local distances distribution of hospitals from the center of their cluster. The 95th percentile threshold is highlighted by the vertical dotted line.

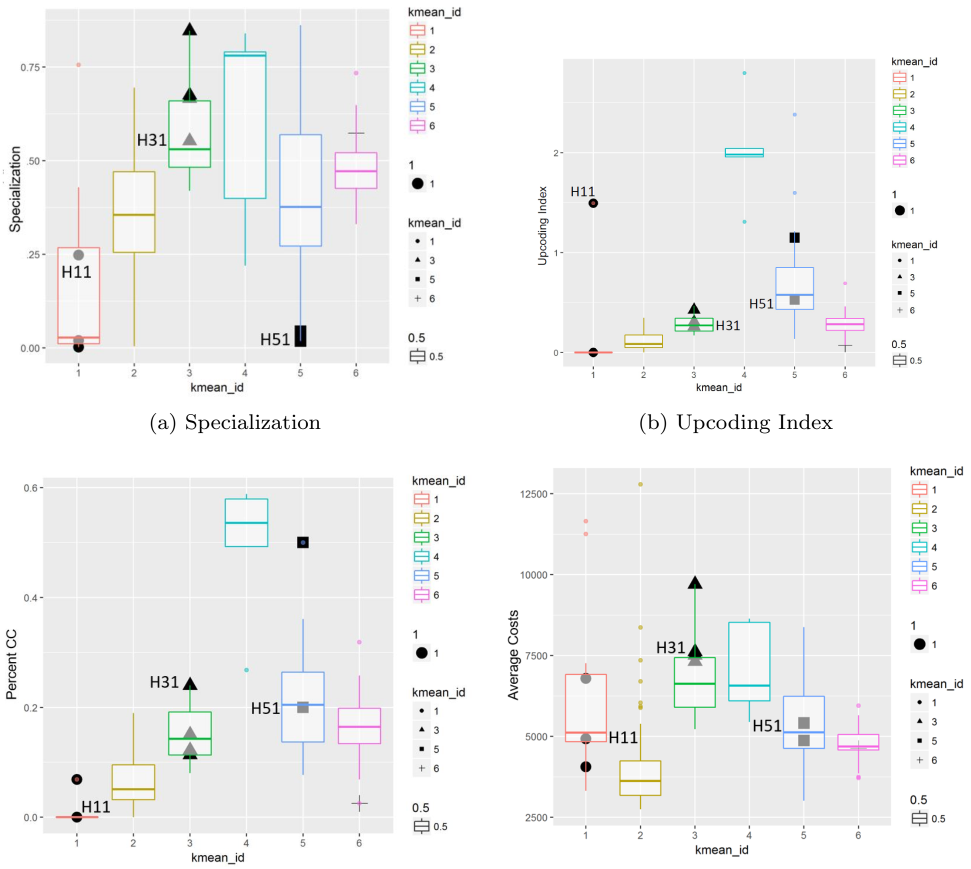
patient's complexity) both aggregated at the patients' level, and for the single SDO, comparing outliers' values with the mean of the whole population.The t-test, as the visual validation, is used to verify whether the two distributions (single patient and entire population of patients) coincide, and if the mean values are actually similar.
Finally, Figure 6 shows the distribution of costs by level of complexity of the case: in this case, the complexity is represented by the length of stay, but similar results were obtained for comorbidity levels.

		Sample Mean			
		Hospital	**All**	**t**	**p-value**
H11	patients' total comorbidity	5.909		0.537	0.5925
H31	patients' total comorbidity	7.258	5.6334	17.024	2.20E-16
H51	patients' total comorbidity	9.895		4.679	6.62E-06
H11	SDO com. weight	0.538		-15.683	2.20E-16
H31	SDO com. weight	2.408	1.879	31.751	2.20E-16
H51	SDO com. weight	2.436		3.3903	8.29E-04

Table 3: T-tests for comorbidity values (patients and SDOs)

In Table 4 are listed some values of reference for the three hospitals.

Hospital H31. This *public* provider has a very high specialization reported in the first dashboard (Figure 4.a), together with a not extremely elevated value of *upcoding index* - even if still relevantly high in respect to the cluster it belongs to (Figure 4.b). Average costs and the percentage of *complications* are high as well (Figure 4.c and 4.d). All together it may suggest that this hospital's results have been largely driven by its strong specialization, but further analysis need to be performed in order to confirm this argument.
Actually, once performed the cross-validation with the patients' population, this assumption



(a) Specialization (b) Upcoding Index

(c) Percentage of CC (d) Average Cost

Fig. 4: Hospitals' variables of interest for outliers' validation

	H11	**H51**	**H31**
Average Age	76.72 years	69.64 years	69.79 years
Average LOS	11.09 days	9.73 days	14.04 days
Average Comorbidity	0.5384	2.4358	2.4081
Total Costs	EUR 576'252	EUR 1'180'253	EUR 75'765'394
Cost/LOS	444.115	556.424	534.667
Cost/Comorbidity	9146.857	2222.699	3117.276
Upcoding Index	1.4933	0.5290	0.2598
Specialization	0.2478	0.0183	0.5524

Table 4: Summary of estimated variables for the three outliers

seems to hold. The patients treated by this provider are on average more complex, according to the proxies adopted by the model. Despite the lower age average (Table 4), the LOS is slightly higher on average (Figure 5), and the comorbidities faced and registered appear significantly above average (Table 3). On top of that, considering the ratios of the reimbursements received for each complexity level (Figure 6), this actor does not demonstrate any particular earning

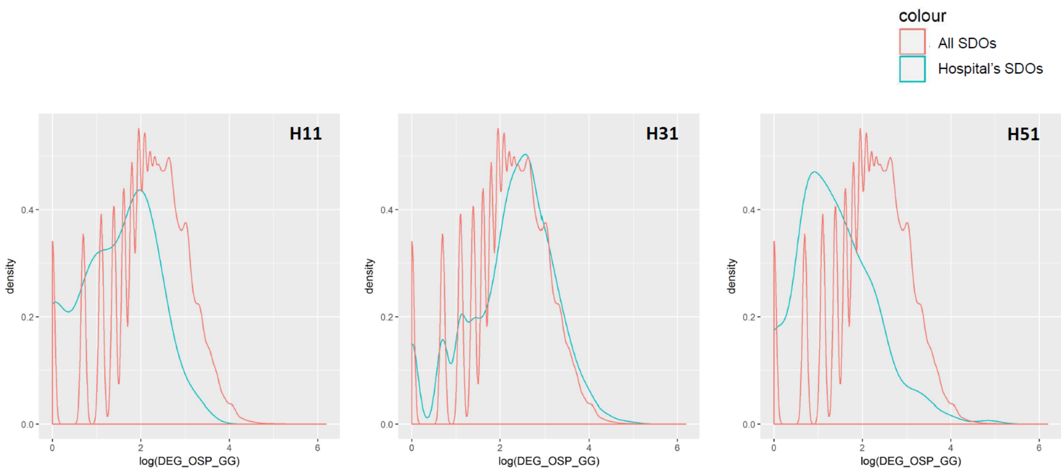


Fig. 5: Logarithm of *Length of Stay* distribution for each hospitals' SDOs compared to all SDOs

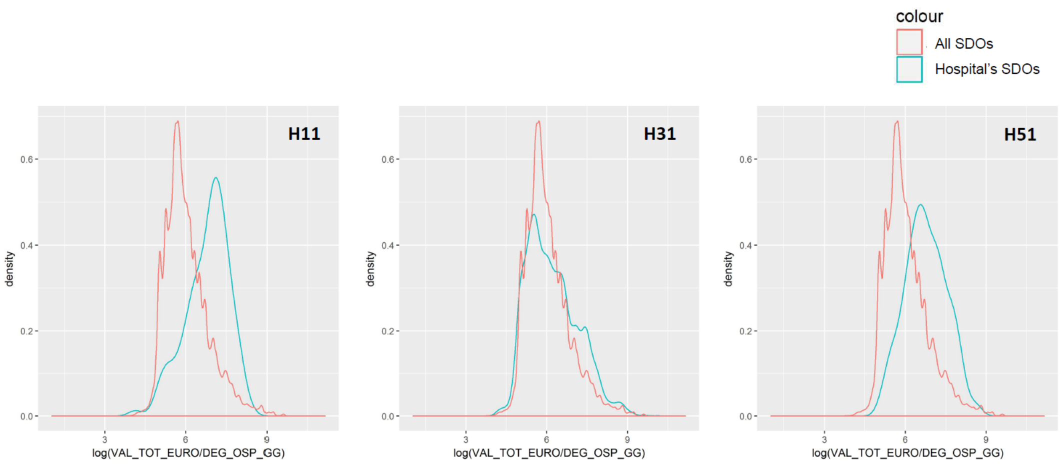


Fig. 6: Logarithm of *Cost/length of Stay* distribution for each hospitals' SDOs compared to all SDOs in the dataset

above the overall group of providers. For this reason, this hospital might be eligible for being cleared with no further in-depth investigation.

Hospital H11 and H51. These two private facilities exhibit opposite results compared to the previously mentioned public provider. Neither of them present high levels of specialization (Figure 4.a), and their upcoding index appears high compared to their cluster and to the population as a whole (Figure 4.b). When cross-validating with their patients' population of reference, results seem to confirm their inappropriate behavior. H11 has both low levels of LOS and comorbidities (Figure 5 and Table 3), while H51 reports a relatively high (or similar to the average of the group) value of co-

morbidity (Table 3), but its patients seem to be treated for shorter periods of time (Figure 5). This may be representative of the fact that the population treated by that hospital has severe comorbidities linked with other pathologies, but for what concerns the treatments received in the time frame under consideration, no long periods of stay were registered. This makes the provider still suspicious. This consideration is supported for both providers when analyzing the reimbursements associated with the complexity of their patients, which appear higher than the average (Figure 6). Even though the two cases are different and may be grounded on different causes, both providers seem to need a more precise in-depth investigation, by someone who may add its domain specific expertise to decide whether an actual fraudulent behavior is in place.
The three cases considered are just a subset of the identified outliers, which were selected for illustrative purposes. It is important to underline once again, that none of the conclusion drawn from this last passage of the model are in any way a definitive judgment of the existence or not of a fraudulent behavior. The aim of this last step is to provide the controllers and the decision makers with a useful set of information in order to ease their appraisal process, and reduce the risk of wasting their effort on false positive cases.

5 Discussion and Conclusions

The objective of this study was proposing a novel methodology to support the preliminary screening phases performed by auditors in the fight against healthcare fraud. It focused on the upcoding practice because of the impact on the system demonstrated by the high attention posed on this topic by researchers in the field (e.g. [2, 27, 43]). The methodology here proposed was tested on real data concerning a complex chronic disease like HF, to verify whether significant insights might be inferred by the obtained results. Theoretically, the proposed model should be promising, although one limit of the research setting is the missing validation of results by domain experts. Another limit that deserves to be highlighted regards the time span of the available data adopted to prove the model here proposed: a period longer than three years might have allowed, for instance, for a better profilation of patients.
A part from the numerical insights commented in Section 4, the outcome of this research deserves some additional remarks.
First and foremost, it should be highlighted that the proposed model can be repurposed for any disease of interest. Additionally, the model is based on an existing and widely adopted coding mechanism (DRG): as such, it can be easily applied to a large number of databases.
Of particular interest are the managerial and policy making implications of this work. The model is fairly simple and easily automated and, since it detects cautiously-perpetrated frauds, it can be used in any context of the healthcare sector. Along with other systems, it can help by reducing the amount of money wasted and the unfair internal system abuses, particularly in the long run. If implemented and sponsored appropriately, the system could deter the rejection of the best practices. Additionally, along with these functionalities, a side system could regularly compute the *distances* of the hospitals from the center of their clusters, thus providing real-time information about which hospitals are drifting away from the top performers of their group, in the spirit of control charts. This would result in a positive feedback system, capable of preventing - rather than detecting - frauds, while encouraging the achievement of the best performance.
The model proposed in this research presents is novelty in different ways when compared to the current state of the art. First, it is a behavioral model [12], aimed at detecting even the

most "cautious" and least evident fraudulent behaviors, contributing to the limited existing literature on the topic. On top of that, it provides a quantitative and systematic procedure, while most of past contributions focused on describing causes and consequences of fraud in healthcare domain, neglecting the proposition of any approach to reduce its impact.
Future research could build on the model here proposed and extend it in various ways, such as testing different clustering techniques, and including Electronic Health Records (EHR) and clinical data in patients' profilation, to better understand their health status and complexity of care.

References

1. Transparency International, Global Corruption Report 2006: Corruption and Health. Pluto Press, London (2006)
2. Berta P., G. Callea, G. Martini, G. Vittadini, The effects of upcoding, cream skimming and readmissions on the Italian hospitals efficiency: A population-based investigation, *Economic Modelling*, **27**, pp. 812–821 (2010)
3. Healthcare.gov, (2018), [online] https://www.healthcare.gov/glossary/fee-for-service/
4. Simborg DW, DRG creep: a new hospital-acquired disease, *New Engl J Med*, **304(26)**, pp. 16021604, (1981)
5. Silverman E., Skinner J, Medicare upcoding and hospital ownership, *Journal of Health Economics*, **23 (2)**, pp. 369–389 (2004)
6. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM., Measuring diagnoses: ICD code accuracy., *Journal of Health Economics*, **40 (5)**, pp. 1620–1639, (2005)
7. Levaggi R., Montefiori M., Horizontal and Vertical Cream Skimming in the Health Care Market, *SSRN Electronic Journal*, **3**, (2003)
8. Friesner D., Rosenman R., Do hospitals practice cream skimming?, *Health Services Management Research*, **22**, pp. 39–49, (2009)
9. Cheng T.C., Haisken-DeNew J.P., Yong J., *Social Science and Medicine*, **132**, pp. 156–164, (2015)
10. Segen's Medical Dictionary. S.v. "*unbundling*." Retrieved from https://medical-dictionary.thefreedictionary.com/unbundling (2018)
11. OECD, Tackling Wasteful Spending on Health, OECD Publishing, Paris
12. Musal R.M., Two models to investigate Medicare fraud within unsupervised databases, *Expert Systems with Applications*, **37 (12)**, pp. 8628–8633, (2010)
13. Steinbusch P., Oostenbrink B., Zuurbier J., Schaepkens F., The Risk of Upcoding in Casemix Systems: A Comparative Study, *Health Policy*, **81**, pp. 298–299, (2007)
14. R Development Core Team (2009), R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [Online] http://www.R-project.org
15. Lorence D.P., Spink A., Regional Variation in Medical Systems Data: Influences on Upcoding, *Journal of Medical Systems*, **26 (5)**, pp. 369–381, (2002)
16. Gonzlez P., On a policy of transferring public patients to private practice, *Health Economics*, **14 (5)**, pp. 513–527, (2005)
17. Rashidian A., Joudaki H., Vian T., No evidence of the effect of the interventions to combat health care fraud and abuse: a systematic review of literature, *PloS One*, **7 (8)**, (2012)
18. Melberg H.O., Olsen C. B., Pedersen K., Did hospitals respond to changes in weights of Diagnosis Related Groups in Norway between 2006 and 2013?, *Health Policy*, **120 (9)**, pp. 992–1000, (2016)

19. Aral K.D., Güvenir H.A., Sabuncuoğlu I, Akar A.R., A Prescription Fraud Detection Model, *Comput. Methods Prog. Biomed.*, **106 (1)**, pp. 37–46, (2012)
20. Pearson R.A., Murray W., Mettenmeyer T., Finding anomalies in Medicare, *Electronic Journal of Health Informatics*, **1 (1)**, p. 2, (2006)
21. Kumar M., Ghani R., Mei ZS, Data Mining to Predict and Prevent Errors in Health Insurance Claims Processing, *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 65–74, (2010)
22. Francis ., Pepper N., and Strong H., C. Francis and N. Pepper and H. Strong, *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 8291–8294, (2011)
23. Kirlidog M., Cuneyt A., A Fraud Detection Approach with Data Mining in Health Insurance, *Procedia - Social and Behavioral Sciences*, **62**, pp. 989–994, (2012)
24. Yang WS, Hwang SY, A process-mining framework for the detection of healthcare fraud and abuse, *Expert Systems with Applications*, **31 (1)**, pp. 56–68, (2006)
25. Shan Y, Jeacocke D., Murray D.W., Sutinen A., Mining Medical Specialist Billing Patterns for Health Service Management, *Proceedings of the 7th Australasian Data Mining Conference*, **87**, pp. 105–110, (2008)
26. Shan Y., Murray D.W., Sutinen A., Discovering Inappropriate Billings with Local Density Based Outlier Detection Method, *Proceedings of the Eighth Australasian Data Mining Conference*, **101**, pp. 93–98, (2009)
27. Luo W., Gallagher M., Unsupervised DRG Upcoding Detection in Healthcare Databases, *2010 IEEE International Conference on Data Mining Workshops*, pp. 600–605, (2010)
28. Joudaki H., Rashidian A. *et al.*, Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature, *Global Journal of Health Science*, **101**, **7**, p. 378, (2015)
29. Tang M., Mendis B., Sumudu U., Murray D.W., Hu Y., Sutinen A., Unsupervised Fraud Detection in Medicare Australia, *Proceedings of the Ninth Australasian Data Mining Conference*, **121**, pp. 103–110, (2011)
30. Konijin R.M., Kowalczyk W., Hunting for Fraudsters in Random Forests, *Hybrid Artificial Intelligent Systems: 7th International Conference - Proceedings*, **Part 1**, pp. 174–185, (2012)
31. Shin H., Park H., Lee J., Jhee W.C., A Scoring Model to Detect Abusive Billing Patterns in Health Insurance Claims, *Expert Systems with Applications*, **39 (8)**, pp. 7441–7450, (2012)
32. Liu Q., Vasarhelyi M., Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information, *29th World Continuous Auditing and Reporting Symposium*, **Part 1**, (2013)
33. Konijin R.M., Duivesteijn W., Kowalczyk W., Knobbe A., Discovering Local Subgroups, with an Application to Fraud Detection, *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference - Proceedings*, **Part 1**, pp. 1–12, (2013)
34. Konijin R.M., Duivesteijn W., Meeng M., Knobbe A., Cost-based quality measures in subgroup discovery, *Journal of Intelligent Information Systems*, **45 (3)**, pp. 337–355, (2015)
35. Van Capelleveen G., Mannes P., Roland M., Dallas T., Van Hillegersberg J., Outlier detection in healthcare fraud: A case study in the Medicaid dental domain, *International Journal of Accounting Information Systems*, **21**, pp. 1831, (2016)
36. Ngufor C., Woytusiak J., Unsupervised labeling of data for supervised learning and its application to medical claims prediction, *Computer Science*, **14 (2)**, (2013)
37. Kose I., Gokturk M., Kilic K., An Interactive Machine-learning-based Electronic Fraud and Abuse Detection System in Healthcare Insurance, *Appl. Soft Comput.*, **36 (C)**, pp. 283–299, (2015)
38. Bolton R.J., Hand D.J., Statistical Fraud Detection: A Review, *Statistical Science*, **17 (3)**, pp. 235–255, (2002)

39. Gagne J.J., Glynn R.J., Avorn J., Levin R., Schneeweiss S., A combined comorbidity score predicted mortality in elderly patients better than existing scores, *Journal of clinical epidemiology*, **64 (7)**, pp. 749–759, (2011)
40. Ekin T., Ieva F., Ruggeri F., Soyer R., On the Use of the Concentration Function in Medical Fraud Assessment, *The American Statistician*, **71 (3)**, pp. 236–241, (2017)
41. Nijssen D., Rousseau R., Van Hecke P., The Lorenz curve: A graphical representation of evenness., *Coenoses*, **13**, pp. 33–38, (1998)
42. Hastie T., Tibshirani R., Friedman J., The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Second Edition, pp. 460–463 , Springer-Verlang, New York (2009)
43. Bauder R., Khoshgoftaar T.M., Seliya N., A survey on the state of healthcare upcoding fraud analysis and detection, *Health Services and Outcomes Research Methodology*, **17 (1)**, pp. 31–55, (2017)

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

46/2018 Riccobelli, D.; Ciarletta, P.
Morpho-elastic model of the tortuous tumour vessels

47/2018 Stefanucci, M.; Sangalli, L.M.; Brutti, P.
PCA-based discrimination of partially observed functional data, with an application to Aneurisk65 dataset

48/2018 Arnone, E.; Azzimonti, L.; Nobile, F.; Sangalli, L.M.
Modeling spatially dependent functional data via regression with di fferential regularization

44/2018 Bernardi, M.S.; Sangalli, L.M.
Modelling spatially dependent functional data by spatial regression with differential regularization

45/2018 Bernardi, M.S.; Carey, M.; Ramsay, J.O.; Sangalli, L.M.
Modeling spatial anisotropy via regression with partial di fferential regularization

43/2018 Fontana, L.; Masci, C.; Ieva, F.; Paganoni, A.M.
Performing Learning Analytics via Generalized Mixed-Effects Trees

42/2018 Antonietti, P.F.; Melas, L.
Algebraic multigrid schemes for high-order discontinuous Galerkin methods

41/2018 Mazzieri, I.; Melas, L.; Smerzini, C.; Stupazzini, M.
The role of near-field ground motion on seismic risk assessment in large urban areas

39/2018 Ferro, N.; Micheletti, S.; Perotto, S.
Density-based inverse homogenization with anisotropically adapted elements

40/2018 Chiappa, A.S.; Micheletti, S.; Peli, R.; Perotto, S.
Mesh adaptation-aided image segmentation