Health Serv Outcomes Res Method (2017) 17:31–55
DOI 10.1007/s10742-016-0154-8

A survey on the state of healthcare upcoding fraud analysis and detection

Richard Bauder[1] · **Taghi M. Khoshgoftaar**[1] · **Naeem Seliya**[2]

Received: 9 November 2015 / Revised: 6 March 2016 / Accepted: 21 July 2016 /
Published online: 28 July 2016

Abstract From its infancy in the 1910s, healthcare group insurance continues to increase, creating a consistently rising burden on the government and taxpayers. The growing number of people enrolled in healthcare programs such as Medicare, along with the enormous volume of money in the healthcare industry, increases the appeal for and risk of fraudulent activities. One such fraud, known as upcoding, is a means by which a provider can obtain additional reimbursement by coding a certain provided service as a more expensive service than what was actually performed. With the proliferation of data mining techniques and the recent and continued availability of public healthcare data, the application of these techniques towards fraud detection, using this increasing cache of data, has the potential to greatly reduce healthcare costs through a more robust detection of upcoding fraud. Presently, there is a sizable body of healthcare fraud detection research available but upcoding fraud studies are limited. Audit data can be difficult to obtain, limiting the usefulness of supervised learning; therefore, other data mining techniques, such as unsupervised learning, must be explored using mostly unlabeled records in order to detect upcoding fraud. This paper is specific to reviewing upcoding fraud analysis and detection research providing an overview of healthcare, upcoding, and a review of the current data mining techniques used therein.

Keywords Healthcare · Healthcare coding · Upcoding · Fraud and abuse · Medicare · Data mining

✉ Richard Bauder
rbauder2014@fau.edu

Taghi M. Khoshgoftaar
khoshgof@fau.edu

Naeem Seliya
nseliya@gmail.com

[1] College of Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL, USA

[2] Health Safety Technologies, LLC, Monticello, NY, USA

1 Introduction

Since the inception of modern group insurance for healthcare in the United States during the 1910's, it continues to be a huge growth industry providing valuable services to an increasing number of people. In general, healthcare costs continue to rise with waste and abuse reduction efforts doing little to lessen these costs.[1] To provide an example for the overall costs involved in US healthcare, the United Nations, the International Monetary Fund, the World Bank, and the Central Intelligence Agency World Factbook[2] indicate the US's annual healthcare spending is larger than the gross domestic product (GDP) of Germany, which was the 4th largest GDP in 2014.[3]

In particular, the rising elderly population, e.g. the baby boomer generation, requires increased healthcare and therefore, appropriate insurance coverage for various medical drugs and services. Medicare is one such insurance growth area for the elderly. Medicare is a government program providing insurance to people over 65 years of age or certain younger individuals with specific medical conditions and disabilities.[4] In general, healthcare payments are made either via the Prospective Payment System (PPS) or Fee-for-Service (FFS) processes, utilizing government funded or private insurance plans. As examples of these payment options, PPS is represented in the Original Medicare Plan which is offered by the federal government, while FFS can be seen in the Medicare private FFS plans for those providers, including physicians, other practitioners, and suppliers that allow this type of reimbursement.

PPS is a predetermined, fixed reimbursement scheme for Medicare payments while FFS is a payment model where doctors and other healthcare providers receive a fee for each service such as an office visit, test, procedure, or other healthcare service.[5] Note that PPS tends to be a more cost-driven healthcare provision, while FFS is more tied to the quality of care. From the National Health Expenditures 2013 Highlights[6] released by the Centers for Medicare and Medicaid Services (CMS), US healthcare spending in 2013 increased 3.6 % to reach $2.9 trillion. Medicare spending alone represented 20 % of all national healthcare spending at about $587 billion, an increase of 3.4 % from 2012. With the increases in healthcare insurance utilization, population growth, the inherent complexity of these programs, and the huge volumes of money involved, this area has been and continues to be attractive for fraud and abuse activities.

1.1 Defining healthcare fraud

The Center for Medicare and Medicaid Services, in their informational pamphlet, Medicare Fraud & Abuse: Prevention, Detection, and Reporting, describe fraud as follows:[7]

[1] http://www.aetna.com/health-reform-connection/aetnas-vision/facts-about-costs.html.

[2] https://www.cia.gov/library/publications/the-world-factbook/index.html.

[3] http://data.worldbank.org/indicator/NY.GDP.MKTP.CD.

[4] https://www.medicare.gov.

[5] https://www.opm.gov/healthcare-insurance/insurance-glossary/.

[6] https://www.cms.gov/Research-Statistics-Data-and-systems/Statistics-Trends-and-reports/NationalHealthExpendData/index.html.

[7] https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/fraud_and_abuse.pdf.

- Knowingly submitting false statements or making misrepresentations of fact to obtain a federal healthcare payment for which no entitlement would otherwise exist.
- Knowingly soliciting, paying, and/or accepting remuneration to induce or reward referrals for items or services reimbursed by federal healthcare programs.
- Making prohibited referrals for certain designated health services.

To further emphasize the importance of detecting and stopping fraud, CMS is clear to point out that "defrauding the Federal government and its programs is illegal. Committing Medicare fraud exposes individuals or entities to potential criminal and civil remedies, including imprisonment, fines, and penalties."

1.2 Impacts of healthcare fraud

With these definitions and concerns in mind, the estimated losses from Medicare fraud alone are in the billions of dollars per year. The Government Accountability Office (GAO) estimated 51 million Medicare beneficiaries in 2013 with services costing upwards of $604 billion. These service costs included nearly $50 billion in improper payments, including documentation and recording errors, within which there may be fraudulent cases (King 2014).

Fraud is generally difficult to assess for insurance programs such as Medicare. The CMS has no reliable way to measure the impacts of fraud, though there are several instruments in place that can help reduce fraudulent activities (Steinbusch et al. 2007). Another critical factor contributing to the difficulty in assessing and detecting fraud is the lack of available and current fraudulent labeled records, which come mostly from insurance carriers, either private or governmental health departments. In order to create labeled records that indicate fraudulent activities, a subject matter expert will audit claims on a limited number of services data. Typically, the auditor re-codes the original medical claim and then compares this re-coded information to the original provider's provided claim (Luo and Gallagher 2010). The auditing processes are governed by guidelines from the GAO[8] with associated privacy protection and a recommendation that audits be performed annually.[9] This implies that audit data can be scarce or limited and be too old for use in current fraud detection schemes. Given this manual detection process, the Federal Bureau of Investigations (FBI) estimates that fraud accounts for 3–10 % of all billings (Morris 2005). With $604 billion in costs during 2013 (King 2014), fraudulent billings could account for anywhere from $18 to $61 billion. Clearly, healthcare fraud continues to be a major problem for the government and taxpayers. The continuing adoption of Electronic Health Records (EHRs) and the subsequent availability of these healthcare records for public consumption, enables data analytics and data mining for fraud and abuse mitigation.

The rapidly increasing use of EHRs continues to drive the collection of massive amounts of healthcare-related data, which has created a demand for big data analytics solutions in healthcare information systems. IDC Health Insights[10] claims by 2018, more than 50 % of big data issues will be handled by routine operational methods, which bodes well for continued fraud mitigation. The use of such large-scale data repositories and the smart application of data science (including data mining and machine learning activities) to

[8] http://www.gao.gov/products/GAO-12-331G.

[9] https://www.aan.com/uploadedFiles/Website_Library_Assets/Documents/3.Practice_Management/1.Reimbursement/1.Billing_and_Coding/7.Audits_RAC/internal.pdf.

[10] http://www.businesswire.com/news/home/20141120005148/en/IDC-Reveals-Health-Insights-Predictions-2015.

detect fraud can lead to substantial cost recovery. For example, it is estimated that with Medicare alone, recovery for 10 or 15 % of expenses through fraud detection is possible (Munro 2014). Moreover, given that the traditional methods for detecting healthcare fraud are inefficient, the motivation for detecting and effectively mitigating these fraudulent activities has far-reaching financial implications for the government and taxpayers.

1.3 Upcoding

One of the most commonly seen fraudulent behaviors in healthcare coding and billing is called upcoding (Swanson 2012). Upcoding is billing for a more expensive service than the one actually performed. It occurs when physicians or medical claims staff enter codes that indicate either services not received or services not fully rendered. For example, office visits can be coded for 30-min visits when only a 10-min visit was actually provided. A related fraudulent activity is undercoding, where providers leave out procedure codes or use codes for relatively lower cost treatments. A provider might undercode a number of services in an effort to avoid audits or to minimize a patient's costs, perhaps as a favor for a friend or relative. Both of these activities, as defined by CMS, are illegal and can incur serious penalties. Upcoding can result in providers charging for unnecessary or expensive services to increase reimbursements. Steinbusch et al. (2007) describe medical claims and reimbursement processes for the US, Australia, and Denmark in great detail and discuss possible factors influencing upcoding by comparing the systems in these different countries. These factors are classified under the casemix system, such as the Diagnosis-Related Group (DRG) classification, and market characteristics. The market characteristics include the size and financial situation of the hospital, whereas the casemix system characteristics are items such as the incentive of the provider, the possibility of changing the coding after the initial registration, and ambiguity of classification criteria used for medical coding.

The focus and intent of this survey paper are to study and investigate the works completed in healthcare fraud, specifically *upcoding fraud analysis and detection*, to help mitigate fraud and reduce financial losses in the healthcare domain. This paper generally does not cover fraud in other domains, such as the insurance or credit card industry, or other healthcare fraudulent activities, like accepting kickbacks for patient referrals. The conclusion of this survey is that upcoding fraud continues to be a serious concern and needs viable solutions via focused innovative research and their validated research solutions. To provide further context, several hindrances to innovative work in upcoding fraud detection, listed below, are noted in the studies investigated and discussed throughout this paper.

- *Data formatting* Hospitals and other healthcare providers create and store data differently with unique formats, variables, and semantics, thus the ability to sort through these differences in order to share and use data is pivotal in providing broader analysis and prediction results.
- *Data sharing* Healthcare organizations can be unwilling to share data due to privacy concerns. Moreover, most patients, for reasons such as privacy or security, do not wish to disclose their healthcare information.
- *Data integration* Throughout the current literature, research is done with either single data sources or by generating new datasets using simple data integration with common features from clearly related data sources. The use of heterogeneous datasets that are loosely related, or do not have common features, can potentially add valuable information overlooked in current homogeneous data sources.

- *Detection methods* Methods currently used for upcoding detection are primarily supervised learning or descriptive statistics and data visualizations. Incorporating different methods and models, tailored to the data, are likely to produce better results in detecting upcoding fraud.
- *Ambiguity* The description of the seriousness of medical procedures, which affects the rate of reimbursement, can be interpreted in different ways. This can lead to difficulties in determining upcoding cases versus normally described procedures.

The remainder of this paper is organized as follows: Sect. 2 provides a brief overview of data mining techniques used in healthcare fraud research. Section 3 presents related works to include other survey papers assessing healthcare fraud detection and the application of machine learning in the healthcare domain, as well as currently used fraud detection applications. Section 4 provides a review of the literature specific to upcoding fraud detection. In Sect. 5, we provide critiques and commentary on the research thus far and what else can be done to attack the upcoding problem. Finally, Sect. 6 presents the conclusions and future work.

2 Data mining techniques in healthcare fraud detection

The purpose of this survey paper is to review current healthcare-related upcoding detection literature; therefore, it is worthwhile to include a brief description of data mining techniques. Data mining, which is often synonymous with machine learning, is a means of extracting pertinent information from data. This data varies and comes from distinct heterogeneous or homogeneous sources and can be structured or unstructured. Structured data is commonly represented in tabular format, and can be stored in an organized fashion in a traditional database allowing the data to be easily processed for data mining, whether homogeneous or heterogeneous in nature. Unstructured data is not obviously organized and requires additional methods of parsing and analysis to apply most machine learning algorithms. For example, text descriptions in medical records are notes annotated by physicians that are usually not formally structured, thus must be appropriately handled in order to extract the underlying structure for the appropriate data mining algorithms. There are three general categories of machine learning as seen in the existing literature: supervised, unsupervised, and hybrid (Witten and Frank 2005).

2.1 Supervised learning

Supervised learning is a popular data mining technique that consists of a target variable (dependent variable or feature) that is used for either prediction or classification from a given set of predictors (independent variables or attributes). Some examples of supervised learners are linear regression, decision trees, random forest, logistic regression, naïve Bayes, and support vector machines. However, this method of learning requires known predictors or class labels in order to model the data for prediction or classification. The requirement for labeled data can be prohibitive with supervised models since this data may either be unavailable or restricted due to legal or privacy concerns. Audit data is typically used as labeled data for healthcare fraud detection but may require an agreement between parties, typically outlined in a non-disclosure agreement. Additionally, the process is limited in the amount of data available and periods of time at which the audits occurred. Publicly available data, such as that provided via CMS, rarely includes any fraud-related

labeled data. Given these difficulties, unsupervised or hybrid learning can be a good alternative.

2.2 Unsupervised learning

Unsupervised learning assumes no apriori target variables (or attributes of interest) to predict or classify. It is used for finding patterns in data by organizing or segmenting a population of records into different groups for further investigation or data label estimation. Some examples of unsupervised learning are association rules, e.g. the Apriori algorithm, and clustering algorithms, such as k-means or mean shift. Unsupervised methods can entail additional uncertainty since the actual links between the measured attributes of the data are unknown, containing no known labels for data records. This is commonly assessed by having a subject matter expert review the groupings to check for correctness. Even with this weakness, unsupervised learning is a valuable tool and can be paired with supervised learning to enable the use of the strong predictive and classification powers of supervised learners with unlabeled data sources via the groupings provided by unsupervised learning. This appears to be a promising technique given most publicly available healthcare datasets are unlabeled.

2.3 Hybrid learning

Hybrid learning methods involve combining supervised and unsupervised algorithms to improve overall prediction or classification results. These methods can leverage the benefits of both types of machine learners applying each learner to specific types of data or at certain times during a learning process. Hybrid methods can use completely unlabeled data records or records containing some labeled data with the rest being unlabeled, which is known as semi-supervised learning. For instance, a clustering algorithm, such as mean shift clustering, can be used to group or partition an unlabeled dataset in order to assess and assign plausible labels for the data record. This newly labeled data are then fed into a supervised learning algorithm in order to predict a class. With regards to applying this technique to healthcare data, the hybrid learning methodology can be used to mitigate the risks associated with either supervised or unsupervised learning alone due to the lack of readily available labeled data as well as the inherent risk of making mistakes when creating new labels.

The three aforementioned machine learning categories are prevalent in healthcare fraud detection literature, but there are other more general data mining techniques used in other industries. These possible methods from other domains, such as the credit card industry, can be leveraged for healthcare fraud detection. Statistical methods, that employ distribution-based models comparing patient or provider profiles, are also seen in the healthcare-related literature. The use of databases and Structured Query Language (SQL) queries written by domain experts can likewise be viewed as a way to mine data in order to assess fraudulent behaviors. Beyond these approaches, data wrangling, feature selection, handling high dimensionality, and big data analytics are all viable data mining techniques for continuing research in upcoding fraud detection along with the methods described in the machine learning categories. For additional information on data mining algorithms, there are many instructive resources to include (Witten and Frank 2005; Tomar and Agarwal 2013; Dave and Dadhich 2013; Gera and Joshi 2015; Ahmad et al. 2015).

3 Related works

The works that relate to our healthcare upcoding fraud analysis and detection survey study include other relevant survey papers since 2010, for which we give a comprehensive review. The review of these recent fraud-related healthcare survey papers examines detection techniques for general fraudulent behaviors and demonstrates a gap in the existing research related to upcoding fraud detection. In addition, we examine two available applications currently used for fraud detection and prevention. The value of these survey papers is not only for comparative purposes but, more importantly, to provide information on what has been used in healthcare fraud detection in order to leverage this collective knowledge for continued upcoding analysis and detection. The following discusses each relevant survey paper and application that we examined.

3.1 Survey studies

Travaille et al. (2011) discusses Medicaid fraud detection while surveying methods in different fields from healthcare to telecommunications to credit cards using several machine learning techniques. The details from other domains are provided in order to determine the applicability of these methods to Medicaid. The authors summarize two Medicaid-related fraud studies clearly focused on detection. In one reviewed study, Furlan and Bajec (2008) describe a general, holistic framework for fraud detection and posit that labeled data is difficult to obtain in the medical field relative to other domains such as the credit card industry. In another reviewed paper, Major and Riedinger (2002) use private insurance claims data for a pilot fraud detection system in six metropolitan areas.

To reassert these data concerns in healthcare, specifically Medicaid, Travaille et al. (2011) note that unlike in other domains, due to the diversity in healthcare data, labeled data is not readily available. Therefore, the use of supervised learning is limited, and unsupervised learning is a better option for a machine learning technique. Even so, unsupervised learning can be prone to false positives, increasing the difficulty in finding legitimate outliers in the records. Besides this limitation, the authors mention that further studies should be done evaluating contemporary industry tools for fraud detection. However, their work provides limited investigation on current studies in Medicaid fraud and is vague in discussing how other domain techniques can be applied to Medicaid healthcare fraud. More details and direct associations between successfully applied techniques in other industries and the discussed issues in Medicaid fraud detection would bolster the usefulness of applying any lessons learned from other industries.

Liu and Vasarhelyi (2013) present a twofold discussion of fraud detection concepts comparing literature on fraud detection techniques using healthcare data in addition to a proposed fraud detection scheme incorporating geo-location data. Survey papers examined by Liu and Vasarhelyi (2013) incorporate supervised learning algorithms, such as multilayer perceptrons (MLP) and decision trees. Additionally, the authors outline a supervised learning process employing genetic algorithms to optimize weights assigned to general practitioner profiles, which are then used by a k-nearest neighbor (kNN) algorithm. Unsupervised techniques reviewed include using self-organizing maps (SOM) to subdivide general practitioner practices and generate a probability distribution model to represent the underlying data and score new cases, with higher scores indicating possible outliers. The authors summarize additional papers that use hybrid approaches applying SOM or clustering to label data records to train MLPs and decision trees.

In Liu and Vasarhelyi's (2013) proposed geo-location fraud detection method, the authors link different insurance claims datasets via unique identifiers and creates groups within this merged dataset. The groups are established by clustering Medicare claims using both the payment amounts and the actual distances between beneficiary and service provider. The authors' geo-location research uses publicly available Medicare inpatient claims data from 2010 only. These works explore machine learning applications primarily to predict for only one kind of fraudulent behavior, such as billing for services not actually performed. Liu and Vasarhelyi (2013) note that supervised methods have incurred more research effort, but labeled data for these methods are not as readily available. The authors suggest that unsupervised techniques, like the geo-location fraud detection method, are generally more applicable.

This study does not provide a thorough or extensive review of existing healthcare fraud detection literature, especially given the broad scope of fraud behaviors described in Section II of the author's paper. There is also no discussion on current industry tools used for fraud detection. Liu and Vasarhelyi (2013) provide a tenuous link between the survey portion of the paper and the geo-location detection method, giving some indication that these topics should be separated into distinct studies. Finally, it is unclear how the author's method compares to other methods described in the survey section of the paper.

In order to outline current challenges in each domain, Phua et al. (2010) explore recently collected literature on automated fraud detection from the past 10 years, including credit card, telecommunications, and healthcare. In addition, the study highlights possible directions taken in other domains to assess interchangeable experiences and knowledge. The authors discuss available structured medical data comprised of demographics, services, and claims. The healthcare-related fraud detection references examined in their paper include a hybrid model where Cox (1995) uses an unsupervised neural network with a neuro-fuzzy classification method. An unsupervised approach by Yamanishi et al. (2004) is also reviewed that employs a SmartSifter algorithm for outlier detection with Australian medical insurance pathology transaction data.

The majority of papers reviewed by Phua et al. (2010) are hybrid or unsupervised methods utilizing unlabeled records from public sources. Interestingly, the authors briefly mention combining training data with evaluation data to be processed by either single or multiple unsupervised learners, hinting at hierarchical models but do not elaborate any further. The authors make note of the fact that much research focuses on complex, non-linear supervised algorithms, when simpler methods, such as naïve Bayes, can produce equal or in some case, better results. Phua et al. (2010) briefly go over other domains but do not clearly link the methods used therein with any of the survey studies or within the broader context of fraud detection.

Ahmad et al. (2015) surveyed healthcare data analysis and fraud detection techniques, highlighting applications and challenges in this domain. The survey areas include several studies specifically discussing fraud detection, but the majority of the research presented is related to diagnosis and disease predictions. The healthcare fraud-related research is limited to one work using supervised learning (Johnson and Nagarur 2015) and two others using unsupervised techniques (Lu and Boritz 2005; Peng et al. 2006). The authors divide their study by data mining techniques including classification, regression, clustering, and association rules. Supervised classification methods discussed include kNN, decision tree, support vector machine (SVM), and naïve Bayes. The healthcare-related classification paper presented takes a multi-stage approach to detect healthcare fraud where Johnson and Nagarur (2015) apply anomaly detection, presumably using both private and public sources. From these anomaly detection results, a risk ranking is created using decision

trees. The first of two unsupervised papers has Lu and Boritz (2005) describing a method to use an adaptive Benfold algorithm to handle incomplete data using health insurance claims data covering general claims for financial reimbursement at Manulife Financial. The claims data used cover a single company's group benefits plan for its employees from 2003 to 2005. In the second referenced unsupervised paper, Peng et al. (2006) apply clustering methods to detect suspicious healthcare fraud activities from large databases. The data used are provided by a US insurance company and contains health claims data from one state; this data is assumed to be private.

Ahmad et al. (2015) note several challenges to healthcare data mining, including the differences in data formats between organizations, quality of data regarding noisy and missing data, and the sharing of data. They discuss several ways to help alleviate some of the described challenges including providing security to ensure data privacy for data sharing and building better medical information systems to accurately capture patient information. Even though the authors describe healthcare data formats and sharing, there are no discussions on aggregating and integrating multiple data sources.

Joudaki et al. (2015) provide an overview of data mining algorithms to identify different approaches to data mining with a focus strictly on healthcare fraud detection. This study is one of the only cited works focused on healthcare fraud, although limited to supervised and unsupervised fraud detection algorithms. Their study discusses several works that demonstrate various applications of machine learning to detect healthcare fraud. One paper, by Liu and Vasarhelyi (2013), incorporates unsupervised learning with clustering for insurance subscriber fraud. In another paper, Kumar et al. (2010) use SVM supervised learning to predict errors in provider insurance claims. Additional relevant studies involve hybrid learning methods incorporating both unsupervised and supervised learning. Ngufor and Wojtusiak (2013) investigate provider fraud, specifically obstetrics claims, using unsupervised data labeling with outlier detection and regression classification. Aral et al. (2012), look at prescription fraud via distance based correlation and risk matrices. The data sources therein range from publicly obtained data from National Health Insurance in Taiwan and Medicare Australia to other data from a presumably private health insurance organization in South Korea.

Joudaki et al. (2015) conclude by recommending seven steps to mining healthcare data, and they are: (1) expert review of attributes; (2) creating and defining new features; (3) identifying outliers; (4) excluding any outliers based on extracted features; (5) investigating these outliers further for fraudulent behaviors; (6) applying supervised learning based on the outlier investigations; and (7) using supervised learning with unsupervised learning to detect new fraud cases and refine the supervised model. Their discussions on data sources are limited with no mention of integrating multiple datasets or the challenges in doing this using common features for linking or combining related but dissimilar data sources.

The survey papers reviewed assess the present and past state of research with regards to general healthcare fraud. Most of these papers have a holistic view on healthcare fraud that does not focus on any particular fraudulent activities, such as upcoding. Additionally, findings of some survey papers, within the healthcare domain, branch into using data mining for disease prediction and readmission studies, not just fraudulent activities. The studies and research referenced by these survey papers include overviews of specific methods and what these methods are aiming to address and detect, but there are limited direct references to upcoding throughout any of the described taxonomies of fraudulent activities. Our work fills this gap by focusing on upcoding fraud and the studies related to analyzing and detecting this unique fraudulent behavior. These reviewed survey papers do,

however, provide a wealth of information pertaining to various algorithmic techniques used in the healthcare domain which can be applied to fraud-related activities, including specific fraud types such as upcoding. The data used in these studies comes from a range of sources that includes private health insurance companies as well as governmental organizations such as Medicare. Discussions on the lack of available labeled data are noted in several of the survey papers and any pertinent references therein, further asserting an unsupervised or hybrid approach as a reasonable method for continued and future fraud and upcoding detection. To the best of our knowledge, this is the first survey paper to provide a thorough examination of articles focused on upcoding fraud analysis and detection.

3.2 Applications

Additional areas of our literature review include software applications used by CMS or other government agencies to detect fraud. The focus on these applications aligns with the purpose of this survey paper in providing information on healthcare fraud detection. Two software tools were found that are currently used for fraud detection and prevention: Opera Solutions software suite[11] and BAE Systems NetReveal.[12] Opera Solutions is used by CMS to help predict fraudulent claims using a mixture of pattern and anomaly detection as well as a risk scoring method for fraud prevention with additional subject matter expert review, essentially combining human and machine for the final assessment of fraud. Opera Solutions also incorporates visualization layers with dashboard capabilities for big data visualizations. The Massachusetts Health Department is using the NetReveal tool. This tool incorporates data ingestion and parsing, predictive and social networking analytics, risk scoring, and other capabilities for fraud detection tailored to specific domains. NetReveal incorporates techniques including unsupervised and supervised learning, rule-based techniques, and network (graph) models. For fraud detection, NetReveal blends outlier detection and peer group analytics to identify providers committing fraud. Both these tools suggest fairly robust implementations but information is limited in terms of actual performance and capabilities. We did not find materials stating or suggesting Opera Solutions or NetReveal's healthcare fraud detection performance and success rates either on the company's websites or via independent reviews.

4 Review of upcoding literature

This section examines upcoding detection and summarizes the research specifically related to upcoding fraud in healthcare. The aforementioned review of related survey papers cover many analyses and algorithms for fraud detection in general, which could be readily applied to upcoding detection, but do not explicitly address the problem of upcoding in healthcare.

4.1 Upcoding detection

One important point to consider prior to the literature descriptions involves the general detection of upcoding. After a patient is admitted, the conditions are assessed and the

[11] http://operasolutions.com/health-exchange-solutions/.

[12] http://www.baesystems.com/product/BAES_166112/netreveal-analytics-platform.

services are provided. The healthcare provider staff (physicians in most cases) annotate this on the patient's medical chart. A medical coder will take the information on these charts and translate them into the appropriate diagnosis (coded as ICD-9 or ICD-10) and procedure (coded as CPT/HCPCS) groups. These diagnoses and groups, when combined with other information, form Diagnosis-Related Groups (DRGs) (Davis 2015).

International Classification of Diseases (ICD) are diagnosis encodings. ICD-9 is currently being phased out, as of October 2015, and being replaced by ICD-10. CPT (Current Procedural Terminology) is a uniform coding system consisting of descriptive terms and identifying codes that are used primarily to identify medical services and procedures furnished by physicians and other health care professionals. CPT contains information concerning the services provided and the intensity of the physicians work, and is used to bill government-provided or private health insurance programs. CPT is the same as a Level I HCPCS (Healthcare Common Procedure Coding System) code for the specific medical service furnished by the provider. Level I is what providers use to report medical procedures and professional services via ambulatory and outpatient settings from physician visits to inpatient procedures. Level II HCPCS codes are used mainly for dental services.

These described diagnoses and procedures make up Diagnosis-Related Groups (DRGs) which are the classification systems that group similar clinical conditions (diagnoses) and the procedures furnished by a hospital during a patient's stay. DRGs are designed so that all patient episodes in a DRG, e.g. mastectomy, consume similar resources minimizing the intra-group variance for a particular provided procedure or "product" to be reimbursed. This is seen as a hierarchical payment process with several points in which upcoding fraud can be injected. Most of the upcoding research focuses on DRGs but others use CPT or provide more generic methods.

4.2 Literature

The existing research works that focus on upcoding are limited and found mostly in health or economics journals with two papers from the IEEE International Conference on Data Mining and ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The date ranges for these related literature papers are from 2000 to 2014. The research is mostly in the area of supervised learning, as was also noted in Sect. 3, with a few other works using or referencing unsupervised methods, statistical and descriptive techniques, text mining, or graph algorithms for fraud detection. The data sources used were a mixture of private and public sources with most using private data with provided audit features. Some CMS data were mentioned but in a limited capacity. Most of these papers directly mention and address the upcoding fraud problem. Due to the limited number of upcoding-specific studies found, we elect to cover each paper in greater detail than other survey papers discussed herein in order to clearly describe the existing body of research and comment on future work.

- *Medicare Skilled Nursing Facility Reimbursement and Upcoding* (Bowblis and Brunt 2014): This work studies the likelihood of upcoding via resource utilization groups (RUG), which are similar to DRGs, in Skilled Nursing Facility (SNF) reimbursements including Medicare payment deltas (or differentials as the authors call them) due to geographic variations, which are akin to cost of living adjustments. RUG codes are composed primarily of the number of therapy minutes and the activities of daily living (ADL) index scores. The ADLs are used in healthcare to assess people's self-care activities which are basic functions, such as getting out of bed, bathing, eating,

grooming, etc. The associated index score is estimated from survey data. For example, over 720 therapy minutes with an ADL index score of 16 indicates a RUG code of RUC that translates into about a $467.21 reimbursement rate. SNFs are places for long term nursing care and services outside of an institution or hospital. Bowblis and Brunt (2014) are assessing if SNFs are using different RUG codes, with varying Medicare reimbursements, across the number of therapy minutes and patient functionality scores due to geographic variations. These changes in RUGs driven by locality can indicate upcoding.

The authors use linear regression to test the geographic reimbursement effect against the SNFs profit margin to show that these cost of living adjustment factors are biased and some areas will have more incentives to upcode because of this geographic variation bias. The regressions report an adjustment variable for a rural facility status (a binary indicator), a wage index variable, and a wage index and rural facility interaction. Generally, rural area SNFs have lower profit margins than urban facilities, and a higher wage index is associated with higher profit margins. The authors claim these results indicate that geographic adjustment factors may provide some areas with more incentive to upcode. Even so, the coefficient of determination, R^2, for all three regression models is below 0.5 possibly indicating a low percentage of variance explained by the regression model. These lower R^2 values are not necessarily poor, but the authors do not expound further on what these R^2 values should be for the model to be valid. In order to model the probability of upcoding, Bowblis and Brunt (2014) use a mixed logit model based on several features including the RUGs geographic differentials, the actual payments received associated with each RUG, and demographics of the SNF and SNF patients. The differentials are the lower and upper differences between a currently chosen RUG code and downcoding or upcoding to a lower or higher reimbursed RUG, respectively. For instance, choosing a RUG code of RUB (which is over 720 therapy minutes with an ADL index score of 13) has a reimbursement rate of $420.56, and raising the RUG code to a higher one of code RUC changes the reimbursement rate to $467.21 with a differential of $467.21 - 420.56 = 46.65$.

Bowblis and Brunt (2014) assess the effects of wage differentials based on therapy minutes and ADLs on the percentage of patients per RUG code. For therapy minutes, they conclude that a higher wage differential incurs a drop in the percentage, or proportion, of patients serviced and an increase in therapy minutes which show possible upcoding. The authors also assess ADLs in relation to upcoding and found no significant difference in patient populations per RUG, concluding that there is no indication of upcoding using ADL. Overall, Bowblis and Brunt (2014) show results that indicate the incentive to upcode was higher due to these geographic payment differentials per RUG selection; ergo, upcoding was found to be more likely in higher cost of living areas. However, their focus is limited to a single multivariate regression model with no comparisons or discussions using other machine learning techniques. There is also limited information on model evaluation statistics, besides listing R^2 and coefficient values, in order to fully assess the models performance, real-world applicability, or detection success rates.

- *CPT Fee Differentials and Visit Upcoding under Medicare Part B* (Brunt 2011): In this study, Brunt uses CPT (HCPCS Level I) codes in researching office visit upcoding in Medicare Part B data. Note that most of the other papers herein use DRGs which are made up of procedure and diagnosis codes, among other variables. The author notes

there can be a big difference between one CPT code and the next code up (similar to the birth weight study for infants and upcoding in Jürges and Köberlein (2013)). The research employs a mixed logit model, like that used in Bowblis and Brunt (2014), applied to physician office visit CPT codes to produce the probability of a particular CPT code for a visit. The data are all publicly available Medicare claims data from CMS, US census information, and income per capita from the Bureau of Economic Analysis.

Out of all features, or explanatory variables, used, the author makes a point to elaborate on two of the features which are the lower and higher fee differentials between office visit CPT codes. These fee differentials indicate the difference between the CPT code a physician should use versus going up or down a CPT code by changing intensity, e.g. length of time of an office visit. Furthermore, as in Bowblis and Brunt (2014), the author acknowledges and accounts for variations in fee differentials based on geographic location and differences in Medicare payments by facility. The geographic location adjustments are based on the Geographic Practice Cost Indices used by Medicare which attempt to adjust provider payments by cost of living.

The specific results reported describe a positive coefficient on the lower fee differential variable which indicates having a lower cost CPT code increases the likelihood of this lower cost CPT code being selected. Conversely, a negative coefficient on the higher fee differential variable suggests the increased cost of the higher CPT code decreases the likelihood of the higher codes selection. The author reports a decrease in the likelihood of choosing the current CPT code by a factor of 0.947 amongst all physicians when the Medicare approval rate for selecting the next higher CPT code increases. In essence, if Medicare is likely to approve the higher fee CPT code, this decreases the likelihood a provider or physician will choose the current lower fee CPT over the higher fee CPT. Overall, Brunt (2011) demonstrates that a large differential increases the likelihood of upcoding services to higher intensity ones and increases the chance of being caught. The author concludes, based on the use of only office visit codes, that physicians have a lot of leeway to adjust the intensity per CPT code; therefore, upcoding is more likely given a larger fee differential.

Brunt (2011) acknowledges the difficulties in estimating the cost of upcoding to Medicare. The author attempts to make an estimate over a subset of physicians with standardized incomes and distribution densities (of physicians) at or below the 25th percentile. The estimate of total cost to Medicare for office visit upcoding is $2.13 billion per year, but this estimate is noted as a rough prediction requiring additional rigor. With regard to the models, the author provides variable and interaction coefficients and significance, along with standard error, but does not provide other model evaluation statistics such as R^2, performance, or detection success criteria or results. Additionally, the use of a single model approach is limited and does not provide any context and comparison for using other machine learning techniques.

- *First Do No Harm-Then Do Not Cheat: DRG Upcoding in German Neonatology* (Jürges and Köberlein 2013): Jürges and Köberlein study the increases in upcoding, from the introduction of DRG usage in Germany, for infant care based on weight and how these weights affect the DRG groupings and hospital reimbursements. The authors indicate that even a small adjustment to a lower recorded weight can increase the reimbursement dramatically by moving the infant care into a different DRG. A voluntary sample from about 13 % of German hospitals was used to compute the per DRG relative cost weight, which is the ratio of resource intensity between different

DRGs, by averaging the treatment costs within each DRG and dividing by the average cost across all DRGs. This ratio represents a higher or lower than average cost assuming 1 to be the average treatment cost per DRG. The main source of data used are official German birth statistics, which include the number of births and birth weights from 1996 to 2010, covering both pre-DRG and post-DRG periods. With neonatal procedures, lower birth weights indicate more resource intensity thus increased reimbursements. Upcoding can then be seen as changing the birth weight to a lower one, thus changing the DRG for increased reimbursement. The authors create birth weight thresholds in order to assess trends in birth weights within each DRG coding. The authors find the percentage of upcoding based on the proportion of births with weights below each threshold, with 50 % meaning no upcoding and a value above 50 % being possible upcoding. For birth weights of 1000, 1250, and 1500 g, they find 80–90 % of recorded birth weights to be below the threshold (recall less birth weight can be in a higher reimbursement DRG) thus a 60–80 % possible upcoding rate based on these thresholds. The authors also estimate the number of upcoding cases per year based on the aforementioned thresholds finding the most upcoded cases being in the 1000, 1500, and 2000 g birth weights. Overall, they estimate over €114 million of excess reimbursements due to neonatal upcoding.

Jürges and Köberlein (2013) use regression to determine relationships between the proportion of upcoded cases and expected payment difference. The conclusions from the regression indicate this relationship is driven almost entirely by the threshold above 1000 g at $p < 0.01$. This confirms the highest frequency and cost of upcoding seen in the 1000, 1500, and 2000 g weight thresholds. There are no notable machine learning techniques applied in this paper, but the authors make good use of descriptive statistics of observed cases to estimate upcoding costs and the number of annual upcoding cases. This is wholly based on finding patterns in existing data and creating differentials from this data to find additional patterns indicative of upcoding. The authors are thorough in their coverage and handling of neonatal birth weight and cost data providing many insights and patterns via graphs and descriptive statistics. Using these patterns and trend data with machine learning could further enhance upcoding detection and predict future cases of upcoding.

- *A Statistical Model to Detect DRG Upcoding* (Rosenberg et al. 2000): This work is a proof of concept study to detect claims upcoding with DRGs. Rosenberg et al. (2000) estimate the probability that a claim has incorrect DRG codes, and with this, provide information to aid in the selection of which claims to investigate and audit based on the predicted upcoding cost recovery versus the cost of the claim audit. This is essentially a method to improve the current detection and audit process. The authors employ supervised learning using audit data with labels created based upon past experiences with DRG upcoding to indicate correct or incorrect DRG codes. The data is assumed to be private provided by Blue Cross Blue Shield of Michigan containing 31,833 audited claims from 1989 to 1998, with 30 % of the DRG codes having been revised, demonstrating an upcoding event.

  The authors use a hierarchical logistic Bayesian model to predict the probability that the DRG is coded incorrectly on the claim based primarily on the principle reason for admission noted at discharge. The principle reason is the primary DRG code that can encompass other DRGs; thus, its use represents the totality of the procedure for admission by the provider. The Bayesian model is composed of a logistic regression model per principle reason that includes the primary DRG feature as well as other features such as age, sex, length of stay, and total paid. The regression parameters are

distributed as multivariate normal with a mean vector θ and variance–covariance matrix. These regression parameters are estimated using a Markov Chain Monte Carlo algorithm with the convergence being determined via trace plots of the mean vector. This method has the advantage of being able to estimate probabilities with a low number of claims or a very low number revised DRGs. The estimate for θ, specifically its posterior distribution, is used to compute the probability that a DRG code is incorrect. A positive coefficient increases the probability a DRG code will need to be revised, i.e. showing possible upcoding.

Rosenberg et al. (2000) run their Bayesian model and two other models with validation and prediction data. The other models used are a standard logistic regression model and what they coin as a 'naïve' model that calculates the expected number of revised DRGs by assuming a static 31.1 % revised DRG rate over all claims, rather than specific revised DRG rates per DRG coding (as provided by the insurer). Tables 1 and 2 provide brief summaries of the validation and prediction test results.

The results on the prediction data clearly show a large difference between observed and expected counts which the authors acknowledge as a point for future research. Even though the other two models appear to outperform the Bayesian model in correct classifications, the authors note this model outperforms the others for the absolute error and squared error measures. The inclusion of other models, such as naïve Bayes and decision tree, should be added for additional performance comparisons. Rosenberg et al. (2000) conclude by using the incorrect DRG code results from the Bayesian model, with the percent of expected recovery and a claim's total paid amount, to estimate the total amount of recoverable claims dollars. With the results from both parts, the probability of incorrect DRGs and estimated recoverable claim dollars, the authors posit that this proof of concept demonstrates auditing 88 % of the indicated upcoded claims would result in recovering 98 % of the overpayments.

- *Medicare Upcoding and Hospital Ownership* (Silverman and Skinner 2004): Silverman and Skinner discuss upcoding in for-profit and not-for-profit hospitals where the upcoding analysis is done using DRGs. This study does not use machine learning techniques but rather descriptive statistics to provide analysis and conclusions, such as regression to find relationships between variables. The authors use publicly available Medicare claims data from 1989 to 1999 and combine this with hospital ownership from the Hospital Association Yearbook by using the Medicare provider number. The primary metric used to evaluate upcoding is the upcoding ratio.

  Silverman and Skinner (2004) focus on DRG and DRG weights for reimbursements for respiratory infections and inflammations with complications, respectively, with the weight being greater for complications thereby creating additional reimbursements. This upcoding ratio is the sum of discharges for respiratory infections and

Table 1 Validation data revised DRG classification

	Count	% of misclassified
Validation data		
# of claims	5278	–
# of true revised DRGs	1671	–
Expected revised DRGs Bayes	1665.7	1.00
Expected revised DRGs logit	1642.3	1.02
Expected revised DRGs 'Naïve'	1640.4	1.02

Table 2 Prediction data revised DRG classification

	Count	% of misclassified
Prediction data		
# of claims	6635	–
# of true revised DRGs	1980	–
Expected revised DRGs Bayes	2143.4	7.62
Expected revised DRGs logit	2124.7	7.31
Expected revised DRGs 'Naïve'	2062.2	4.15

inflammations with complications over the sum of discharges of the four total DRGs, to include respiratory infections and inflammations, that comprise general respiratory ailments with and without complications. With these upcoding ratios, the authors claim the incentive to upcode from the other three DRGs to respiratory infections, which is the higher reimbursement DRG, is an extra $2000 per discharge. They use the upcoding ratio to measure the amount of upcoding based on the interaction between the market and for-profit and not-for-profit hospitals. Some of the conclusions show that for-profit hospitals, with a 25–50 % market share, had a 45.6 % upcoding ratio in 1997, whereas the not-for-profit hospitals (both private and government) had 34.1 and 30.0 % upcoding ratios with the same market share.

As noted, there is no machine learning or predictive component present in this work. Additionally, the use of the upcoding ratio alone does not seem to be a rigorous metric for prediction or classification. This metric used in conjunction with other metrics, such as DRG cost differentials, may provide further insight. The market share results tend to show an inclination to upcode at for-profit hospitals, though the authors' findings are not conclusive. Some other conclusions from their work are that upcoding is due to administrators who direct coders to upcode and physicians filling out records with information that could lead to using a higher DRG claim whether purposefully or unintentionally.

- *Unsupervised DRG Upcoding Detection in Healthcare Databases* (Luo and Gallagher 2010): Luo and Gallagher present research on DRG upcoding detection that used DRGs and claim coding data on a per hospital basis. The authors claim the results from their method are consistent with domain expert assumptions. In order to support this claim, they employ an unsupervised learning method which, interestingly, is the only paper we found specifically targeting upcoding in this manner (with the other studies using supervised learning or descriptive statistics). Instead of audit data, they use hospital data with DRG information specifically for hip replacements (orthopedic) and acute myocardial infarction. The authors use distributions of DRGs between hospitals to detect differences indicating upcoding, with the assumption that most hospitals will have similar DRG distributions, i.e. aberrant distributions versus normal.

  Luo and Gallagher (2010) hypothesize that a difference in distribution can be detected using leave-one-out cross validation, because there is no way to know which hospitals may actually have upcoding, since the data is unlabeled. They apply two steps to create their unsupervised model. First, they use the decision tree in the *rpart* package in R to classify patient cases into homogeneous subgroups based on procedures, diagnoses, and DRGs. The Gini index is used as the impurity measure where a lower value indicates a more pure or homogeneous set of instances. This index indicates when to stop the splitting process either with a very low impurity measure or when no other variable can split the node to decrease the index further.

In this study, Luo and Gallagher (2010) use I03A, I03B, and I03C as DRG split indicators for the hip replacement. The decision tree model predicts the DRG split indicator with each leaf node of the tree containing some number of homogeneous instances. From these subgroups, the authors count the number of instances per leaf node that belong to each DRG split indicator. For the second step, the authors take these counts per indicator, formed as contingency tables, and perform a pairwise comparison over all hospitals using Fischer's exact test for statistical independence. With this test, the rejection of the null hypothesis indicates two different distributions. In one example, Luo and Gallagher (2010) demonstrate that for hip replacement/revision DRG code I03, using 11 public hospitals with 1 year's worth of code and data, there is one hospital found that indicates possible upcoding. They show this hospital may exhibit upcoding because 25 % of all 59 patient cases have a DRG split on B, whereas only 12 % of all 1030 patient cases for the other ten hospitals have a split on the B indicator. Additionally, Fischer's exact test reported a p value of 0.008 therefore rejecting the null hypothesis.

The authors conclude that the application of this unsupervised algorithm, in testing, with the orthopedic and acute myocardial infarction data, seems to demonstrate some model effectiveness. This study shows promise, particularly with subgrouping, but the authors do not assess other methods of partitioning such as clustering or SOM. Further, there are no performance or detection success metrics defined, nor any tests with comparisons to known upcoding cases at these hospitals. Even if there is no audit data available, as the authors suggest, finding and using a case of known, observed DRG upcoding cases for comparison would add critical validation to their proposed solution.

- *Knowledge Discovery from Massive Healthcare Claims Data* (Chandola et al. 2013): This is a general coverage paper that assesses healthcare data via social network analysis, text mining, and temporal analysis. Chandola et al. (2013) use a time stamped dataset which included labels for fraudulent providers from the Texas Office of Inspector General's exclusion database. They discuss the use of typical treatment profiles to compare among providers to spot possible misuses or abuses in procedures to treat particular ailments.

  For text mining on claims and enrollment data with labeled fraudulent cases, Chandola et al. (2013) create a document term matrix of providers versus diagnoses, where the providers are seen as documents and the diagnoses are terms, and other term matrix combinations. The authors show, using Latent Dirichlet Allocation, patterns such as Diabetes and Dermatoses joined together, warranting further investigation to determine legitimacy. The authors also employ Social Network Analysis where graphs are used with a node's features to discriminate between fraudulent and non-fraudulent cases. Using features from the fully constructed graph on a per node basis with methods such as clusters, degree, connectivity, flows, shortest path, link analysis (page ranks, hits), Eigenvectors, and k-clique, the authors propose the data can then be classified based on these outliers from the graph features. Temporal analysis is a technique used in this study to compare claim submittal patterns over time, per provider, to estimate population norms for similar provider types. Features are defined based on these normal versus anomalous patterns, therefore creating the fraudulent activity detection.

  Chandola et al. (2013) describe the results of the aforesaid techniques generally, without necessarily tying together the various techniques and results, thus results are not included. This appears to align with the intent of the study to translate the problem of healthcare data analysis into some well-known data mining methods. Note the inclusion of this paper herein is based on the general discussion on distribution or

profile comparisons to detect fraud. The techniques described in this study are similar to some of the methods described in other upcoding-related papers and lend themselves to possible applications in upcoding detection.

- *Variation in Charges for Emergency Department Visits across California* (Hsia and Antwi 2014): This study describes a process for detection of charge variations to determine whether hospital- or market-level factors influence these charges. Though the authors mention upcoding, this study is not directly linked to upcoding detection, but the data are similar and regression is used to explain charge variability using healthcare and market data similar to other referenced studies. There is no machine learning used for prediction of charges, but the value in looking at this paper, with regard to upcoding detection, stems from the data used, how the data were used, any variable interactions described via the model, and discussions on the explanatory power of the model.

  Hsia and Antwi (2014) investigate variations in cost between CPT procedure codes for hospital emergency room (ER) visits across Levels 2–4 (out of 1–5) indicating ER service intensity. The authors attempt to show how and if this variability can be explained. They use emergency department, or emergency room, charge data from every non-federal California hospital in 2011 from the Office of Statewide Health Planning and Development. Multivariate linear regression models are used to explain the variation in cost between CPTs for ER visits via the coefficient of determination, R^2, measure. The authors state that the model explained 30 % (Level 2), 39 % (Level 3), and 41 % (Level 4) of the hospital charge variation between ER visits. In order to account for both hospital and market information, they include hospital ownership, number of beds, wages averaged over 3 years, and other factors, for hospital influences and wage index, percentage uninsured, percentage below poverty line, and a measure of market concentration. To get these predictor variables, the authors perform limited data linking via common features in the hospital financial and utilization data, census data, and Medicare impact files (relative cost of living and casemix index). They then group and compare the different hospital types: government, for-profit, and not-for-profit.

  Hsia and Antwi (2014) indicate several limitations, including not capturing all the relevant hospital and market variables as well as being unable to account for all cost differentials. They mention upcoding but claim the cost difference due to upcoding is not significant enough to explain these charge variations as noted by the R^2 values. There is no reason, beyond the fact that there are periodic audits per CMS guidelines and HCPCS standards are in widespread use, to justify discarding upcoding as a possible influence on these charge variations. The authors conclude that charges for ER visits are not predictable having a lot of unexplained cost variation, thus adversely affecting uninsured patients and insured patients using out of network care.
- *Detection of Upcoding and Code Gaming Fraud and Abuse in Prospective Payment Healthcare Systems* (Suresh et al. 2014): This document describes a patented fraud detection system. This invention is designed for the detection of fraud and abuse in hierarchical coded payment systems like PPS. The system analyzes, classifies, and predicts at each level of the hierarchical payment system but can also be used with non-hierarchical payments, such as FFS providers. Suresh et al. (2014) claim that any payment context with at least two payment classification levels, e.g. primary and secondary levels within DRGs, can be used as inputs in the system. The authors primarily focus on upcoding detection throughout the patent description document.

  In general, to detect possible upcoding, Suresh et al. (2014) employ an all-possible-

pairs analysis for which each possible pair of groups contains a normal distribution for a certain, specified metric, which the authors refer to as profiling. They use an unspecified unsupervised learner to detect significant departures from the normal amongst pairs of groups. These distribution differences are assessed per level of the hierarchical payment scheme, i.e. profiling across and within category, group, and element levels. A category is the highest, most general level which includes codings such as the Major Diagnostic Category (MDC). MDC splits ICD-9 codes into 25 mutually exclusive areas; alternately, an MDC can also be groupings of DRGs.[13] The group level falls below the category, with an example being the DRG. Finally, the element is the lowest level and includes features such as age, HCPCS I and II codes, comorbidities, etc. These categories, groups, and elements are associated with some overall payment systems like Skilled Nursing Facility PPS.

Suresh et al. (2014) describe the general process for upcoding detection via profiling across or within each classification level. Across classification levels refers to detection over multiple MDCs or DRGs, whereas within classification levels extracts information from within a DRG, for instance. The category level is purported to be a high-level indicator of the nature of work at a facility. The authors provide examples for each level/across and level/within pairing but note these examples are for illustrative purposes, rather than points of model evaluation or performance. The process of looking for aberrant behavior versus normal behavior is applied to all of the level combinations. As an example of within category detection, the authors illustrate aberrant behavior with an MDC for a facility having a weight of 2.25 while the normal weight is 1.25. Typically a grouper software maps these elements into groups to determine payments, but the elements that make up these groups can be the progenitors for upcoding activities. The authors give an example of one of these elements as the principle diagnosis on a claim for inpatient hospital PPS. If a hospital uses only one or two principle diagnoses to compose a DRG when up to 30 can actually be used, and the normal distribution of principle diagnoses per the population is more varied across the 30 elements composing this DRG, this hospital's DRG composition may indicate aberrant behavior requiring further investigation.

We note, however, that there are no discussions by the authors on any limitations of their invention. The patent document also lacks any evaluation or performance statistics or discussions on detection success rates. It does not specify whether the system is in use in the healthcare field or how it performs relative to other upcoding detection methods and applications. The methods implemented in this invention, in general, appear to be in line with anomaly detection methods comparing various distributions against a normal distribution to detect upcoding behavior. This is not the only study herein to have employed unsupervised methods to assess differences in distributions to detect upcoding fraud. The novelty of the Suresh et al. (2014) invention versus the other studies discussed, is the handling and leveraging of the hierarchical nature of the payment and medical coding system. Additional studies and research incorporating the hierarchical nature of the healthcare systems should be explored comparing various machine learning techniques and results.

- *Methods to Detect DRG-Upcoding* (Schönfelder and Klewer 2008): Schönfelder and Klewer discuss the detection of upcoding to reduce the number of checks and audits on real, non-fraudulent cases, which is very similar to Rosenberg et al. (2000). This method recommends a possible instance of upcoding if the cost of the inspection by the

[13] http://health.utah.gov/opha/IBIShelp/codes/MDC.htm.

insurance company is less than the recovered costs from following through with the upcoding investigation and possible prosecution. The authors claim some detected upcoding instances were found to be correctly billed claims (false positives), resulting in wasted resources. Manual controls and audits done by the Medizinischen Dienstes der Krankenversicherung (MDK) in Germany indicate that only 44 % of all selected suspicious cases are incorrectly coded, leading to unnecessary and costly checks on correctly coded claims.

In order to help automate the selection process before being investigated by an organization like the MDK, Schönfelder and Klewer (2008) use the logistic regression model in SPSS 14.0 to calculate the probability of upcoding using 8000 inpatient claims bills. The regression model includes predictor variables such as age, length of hospital stay, number of secondary diagnoses, cost differentials, and revenue, along with variable interactions such as age and the number of secondary diagnoses. The authors run the model with all DRG codes and subgroups of DRG codes, per specific diagnoses. Schönfelder and Klewer (2008) use the cost of an investigation in addition to other associated expenses, such as the investigators and staff. They assume a fixed amount per case for the calculation of recoverable amounts due to an erroneous payment or upcoding to assess whether to pursue a claim investigation. The logistic regression model results, using all DRGs, showed 62.4 % of cases were correctly classified. The results on the DRG subgroups indicate better results, such as an 81 % rate of correct classification for the gastroscopy subgroup and 70 % for respiratory system infection and inflammation. There are, however, three subgroups with results below 20 % classification success. Even with these lower classification success rates, the analysis for DRG subgroups produces a more accurate identification of upcoding cases for use in comparing audit cost versus recovery costs.

Schönfelder and Klewer (2008) do not indicate any model performance, error metrics, or variable and interaction significance. The correctly detected coding probabilities appear useful, but the authors do not state or recommend what threshold should be applied to indicate when to audit or not to audit. Additionally, no other models were used for comparison purposes. We attempted to assess these results against those in Rosenberg et al. (2000), specifically using the predicted data, and found that the latter study appears to show much higher classification success rates. To be fair, this study does not examine observed and predicted upcoding cases and the results are presented differently, even though both use very similar methods.

- *Top Billing: Meet the Docs who Charge Medicare Top Dollar for Office Visits* (Ornstein and Grochowski 2014): This article does not apply machine learning but instead incorporates descriptive statistics to assess office visit fraud. Ornstein and Grochowski (2014) look at office visit CPT codes and calculate variations among providers with Medicare Part B data. They use service and payment information from the 2012 Medicare Part B data with over 440,000 billings for office visits with a minimum of 11 patients. Based on the nature of the office visit, the CPT codes signify the severity of a visit from "1" being a short visit to "5" indicating an intense examination requiring more time. The most common code is "3" being an average office visit. The authors compare billings across peer groups by grouping procedures and providers. Ornstein and Grochowski (2014), using the 99215 zip code only, are able to discover 600 providers who billed at level "5" more than 90 % of the time. Overall, 20,000 providers billed almost exclusively with "4" or "5" level office visits. Ornstein and Grochowski (2014) show that there are cases where the highest severity of office visit CPT code was used with little to no real justification. Some of these cases

are at teaching hospitals where senior doctors would take on complex cases, but most of these higher level CPT codes are not, which could show fraudulent cases and warrant further investigation. This study is limited to one zip code and one type of visit, i.e. a routine office visit. This survey article is essentially a summary of the analysis which consisted of gathering, grouping, and analyzing the data using descriptive statistics. The interest in this paper stems from the use of publicly available Medicare data, the CPT focus, and comparisons with providers. The initial work described in this article is useful due to the discussions on data preprocessing and visualization as a precursor to more detailed analysis and the application of machine learning methods.

5 Discussion

The papers discussed are primarily concerned with upcoding fraud detection and analysis using either machine learning or descriptive statistics and analysis to draw conclusions. Some of the survey papers indicate model evaluation statistics and results for the described experiments. The model evaluations, test results, and detection success rates are limited with regard to upcoding detection. The fact that presentation of model success rates and results are limited demonstrate the overall lack of innovative research in the area of upcoding fraud detection. This section includes some of the more salient commentary and discussion regarding the previously discussed upcoding-related literature.

- There are a limited number of supervised and unsupervised learning techniques applied to healthcare fraud as discussed in Sects. 2 and 3, but this number dwindles for upcoding specific detection. The types of learners employed were linear regression, mixed logit, and hierarchical logistic Bayesian models as supervised techniques and a combination of subgroup creation via decision tree and Fisher's Exact Test for the unsupervised learner. Given these deficiencies in upcoding detection research and application, there is room to research and apply additional learning and classification techniques. The use of logistic regression is prevalent in the upcoding literature along with multivariate linear regression. Other fraud-related literature suggests decision trees are commonly used supervised learning methods, but these are not mentioned in any supervised learning upcoding literature. Additionally, the literature is lacking in research which utilizes ensemble or boosted learners applied to algorithms such as bagging with naïve Bayes or decision tree boosting.
- The only discussions on strictly unsupervised learning were in Luo and Gallagher (2010) and Suresh et al. (2014). Luo and Gallagher (2010) employ distribution comparisons via Fisher's Exact Test, but the authors do not offer any comparisons using other parametric or nonparametric tests for probability distribution equality such as Kolmogorov–Smirnov, Mann–Whitney, or the Student's t test (assuming distribution normality). The grouping described therein is generated with a decision tree, but there was no discussion on validation of these groups or comparisons to other methods such as k-means clustering or hierarchical clustering for this group creation. Suresh et al. (2014), in their patent documentation, develop an unsupervised method for the comparison of distributions in hierarchical payment systems but do not offer any alternatives or evaluation statistics on their technique with test data or versus other methods. Note that it is not known what model is used by Suresh et al. (2014). Given

there is only one clear description of what unsupervised learning method was used, there are opportunities to apply other unsupervised techniques such as transfer learning, active learning, deep learning, or ramp detection algorithms looking for shifts between steady states. The use of hybrid learning techniques is another avenue of research in combining supervised and unsupervised learning. For instance, an unsupervised learner can create data labels that can be used to train a supervised model. Conversely, if labeled data is available, a supervised learner can be created with an unsupervised learning algorithm, such as anomaly detection, used to detect new upcoding behaviors which are used to retrain the existing supervised learning model.

- The use of labeled data obtained through audits is a limiting factor in any healthcare analysis as this type of data is in itself limited in scope, from the amount of data actually labeled by auditors to the time frames available for these audit datasets. Labeled records could also be difficult to obtain due to legal and privacy restrictions. It's hard to say how extensible this type of research would be given the expeditiously increasing volumes of healthcare data and the reliance on standard audit policy and practices. A majority of the upcoding research appears to focus on using labeled data for supervised learning, with the assumption, if not clearly stated, that the unsupervised learning algorithms presented are using unlabeled data. Therefore, the use of unsupervised learning, as discussed, is a reasonable method to employ due to the availability of predominately unlabeled healthcare data.
- The literature contains some discussions on linking different data sources based on common features prior to any data analysis and application of machine learning, but were minimal discussions on the fusion of heterogeneous, or heterogeneous and homogeneous, datasets or the integration of model results from different training datasets. Given the range of data publicly available, it makes senses to try to gather as much relevant information as possible to model and detect upcoding fraud. The limitations of the papers herein indicate using features from a small number of readily available data sources. CMS alone has a panoply of datasets available, most with common features for linking, so even at the simplest form of data integration, different CMS datasets can be linked to provide additional, usable information for upcoding fraud analysis and detection. As mentioned, heterogeneous and seemingly unrelated data can also be brought together to assess patterns prior to or after the use of data mining techniques for upcoding detection.

6 Conclusion

The papers reviewed focused more on the technical algorithmic methods to either analyze or detect upcoding with healthcare data. It has been shown in Sect. 3 that there is a plethora of research for healthcare fraud detection, but the field is sparse when it comes to the specifics of upcoding fraud detection. The healthcare areas of overtreating, overtesting, and overcharging are all ripe for upcoding activities. Most importantly, fraud, including upcoding, is of great concern due to the huge financial impacts. The detection of fraud is critical in reducing the number of fraudulent cases and reducing these financial burdens. The survey papers address the growing financial impacts of fraud and upcoding but do not clearly discuss who benefits from upcoding detection. A large number of current news involves Medicare or Medicare-related upcoding fraud cases (Bricker 2015; Sweeney

2015)[14,15], indicating these programs still need focus in creating and applying effective upcoding detection methods.

The current upcoding research describes mostly supervised learning methods applied largely to labeled records from audited claims data. This is a limiting factor in continued upcoding analysis and detection due to this focus on primarily supervised learning. Therefore, the use of publicly available, mostly unlabeled, data should be pursued smartly, applying data mining and machine learning. The use of unsupervised or hybrid learning is a logical path forward in assessing upcoding fraud. Continued research will involve the use of clustering or the comparison of probability distributions, both of which appear to be promising in detecting upcoding. The patent document (Suresh et al. 2014) asserts the hierarchical nature of healthcare payment systems in modeling and detecting upcoding fraud. There is no reference to the type of unsurpervised learner used, but research into using models leveraging layered information, such as a hierarchical Bayesian model or deep belief network learners, should be further explored in order to best capture the nuances of the healthcare payment and reimbursement processes for upcoding detection.

More in-depth data integration with publicly available big data sources, beyond those summarized in the reviewed literature, can also add to the meaningful detection of this type of fraud by including more relevant information and patterns aiding machine learning techniques. For example, CMS provides massive amounts of data ranging from Medicare claims to physician referrals and hospital utilization, all of which should be assessed for possible data integration and applications of machine learning. Several of these kinds of data sources are mentioned in the survey works, but no study has demonstrated expansive use of available data or extensively described data integration. Additionally, further research is needed in order to incorporate feasible solutions into current systems or create new systems to assist corporate audit departments in expediting audits and detecting fraudulent activities.[16] Other future work will include further exploration applying text mining and network (graph) analysis to create labeled data as well as methods to better understand interactions between providers, especially with many integrated data sources, via referral analysis to mine any additional patterns for the detection of upcoding.

Acknowledgments The authors would like to thank the editor and the anonymous reviewers for their insightful comments. They would also like to thank various members of the Data Mining and Machine Learning Laboratory, Florida Atlantic University, Boca Raton, for their assistance reviewing this manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ahmad, P., Qamar, S., Rizvi, S.Q.A.: Techniques of data mining in healthcare: a review. Int. J. Comput. Appl. **120**(15), 38–50 (2015)

[14] http://health.wusf.usf.edu/post/medicare-plans-upcoding-cost-billions.

[15] http://khn.org/morning-breakout/states-parkland/.

[16] http://www.cigna.com/reportfraud/index.

Aral, K.D., Güvenir, H.A., Sabuncuoğlu, İ., Akar, A.R.: A prescription fraud detection model. Comput. Methods Programs Biomed. **106**(1), 37–46 (2012)
Bowblis, J.R., Brunt, C.S.: Medicare skilled nursing facility reimbursement and upcoding. Health Econ. **23**(7), 821–840 (2014)
Bricker, E.: Physician upcoding: Does it happen? If so, how?. http://www.compassphs.com/blog/uncategorized/physician-upcoding-does-it-happen-if-so-how-2/ (2015)
Brunt, C.S.: CPT fee differentials and visit upcoding under Medicare Part B. Health Econ. **20**(7), 831–841 (2011)
Chandola, V., Sukumar, S.R., Schryver, J. C.: Knowledge discovery from massive healthcare claims data. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '13, pp. 1312–1320. ACM, New York, NY (2013)
Cox, E.: A fuzzy system for detecting anomalous behaviors in healthcare provider claims. In: Intelligent Systems for Finance and Business, pp. 111–134. John Wiley & Sons (1995)
Dave, D.M., Dadhich, P.: Applications of data mining techniques: empowering quality healthcare services. In: IJICCT (2013)
Davis, E.: DRG 101: What is a DRG & How does it work?. http://healthinsurance.about.com/od/medicare/fl/DRG-101-What-Is-a-DRG-amp-How-Does-It-Work.htm (2015)
Furlan, Š., Bajec, M.: Holistic approach to fraud management in health insurance. J. Inf. Organ. Sci. **32**(2), 99–114 (2008)
Gera, C., Joshi, K.: A survey on data mining techniques in the medicative field. Int. J. Comput. Appl. **113**(13), 32–35 (2015)
Hsia, R.Y., Antwi, Y.A.: Variation in charges for emergency department visits across California. Ann. Emerg. Med. **64**(2), 120–126.e4 (2014)
Johnson, M.E., Nagarur, N.: Multi-stage methodology to detect health insurance claim fraud. Health Care Manag. Sci. (2015). doi:10.1007/s10729-015-9317-3
Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., Arab, M.: Using data mining to detect health care fraud and abuse: a review of literature. Glob. J. Health Sci. **7**(1), 194 (2015)
Jürges, H., Köberlein, J.: First do no harm. Then do not cheat: DRG upcoding in German Neonatology. CESifo Group Munich, CESifo Working Paper Series 4341 (2013)
King, K.M.: Medicare fraud: progress made, but more action needed to address Medicare fraud, waste, and abuse. http://www.gao.gov/products/GAO-14-560T (2014)
Kumar, M., Ghani, R., Mei, Z.-S.: Data mining to predict and prevent errors in health insurance claims processing. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '10, pp. 65–74. ACM, New York, NY (2010)
Liu, Q., Vasarhelyi, M.: Healthcare fraud detection: A survey and a clustering model incorporating geo-location information. In: 29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia (2013)
Lu, F., Boritz, J.E.: Detecting fraud in health insurance data: learning to model incomplete Benford's law distributions. In: Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3–7, 2005. Proceedings, pp. 633–640. Springer, Berlin (2005)
Luo, W., Gallagher, M.: Unsupervised DRG upcoding detection in healthcare databases. In: 2010 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 600–605 (2010)
Major, J.A., Riedinger, D.R.: EFD: a hybrid knowledge/statistical-based system for the detection of fraud. J. Risk Insur. **69**(3), 309–324 (2002)
Morris, L.: Combating fraud in health care: an essential component of any cost containment strategy. http://content.healthaffairs.org/content/28/5/1351.full (2009)
Munro, D.: Annual US healthcare spending hits $3.8 trillion. http://www.forbes.com/sites/danmunro/2014/02/02/annual-u-s-healthcare-spending-hits-3-8-trillion/ (2014)
Ngufor, C., Wojtusiak, A.: Unsupervised labeling of data for supervised learning and its application to medical claims prediction. Comput. Sci. **14**(2), 191–214 (2013)
Ornstein, C., Grochowski, R.J.: Top billing: meet the docs who charge Medicare top dollar for office visits. https://www.propublica.org/article/billing-to-the-max-docs-charge-medicare-top-rate-for-office-visits (2014)
Peng, Y., Kou, G., Sabatka, A., Chen, Z., Khazanchi, D., Shi, Y.: Application of clustering methods to health insurance fraud detection. In: 2006 International Conference on Service Systems and Service Management vol. 1, pp. 116–120 (2006)
Phua, C., Lee, V.C.S., Smith-Miles, K., Gayler, R.W.: A comprehensive survey of data mining-based fraud detection research. In: CoRR. arXiv:1009.6119 (2010)

Rosenberg, M.A., Fryback, D.G., Katz, D.A.: A statistical model to detect DRG upcoding. Health Serv. Outcomes Res. Method. **1**(3), 233–252 (2000)

Schönfelder, T., Klewer, J.: Methods to detect DRG-upcoding. Heilberufe **60**, 6–12 (2008)

Silverman, E., Skinner, J.: Medicare upcoding and hospital ownership. J. Health Econ. **23**(2), 369–389 (2004)

Steinbusch, P.J., Oostenbrink, J.B., Zuurbier, J.J., Schaepkens, F.J.: The risk of upcoding in casemix systems: a comparative study. Health Policy **81**(23), 289–299 (2007)

Suresh, N., de Traversay, J., Gollamudi, H., Pathria, A., Tyler, M.: Detection of upcoding and code gaming fraud and abuse in prospective payment healthcare systems. US Patent 8,666,757 (2014)

Swanson, T.: The 5 most common types of medical billing fraud. http://www.business2community.com/health-wellness/the-5-most-common-types-of-medical-billing-fraud-0234197 (2012)

Sweeney, E.: Florida fraud case highlights concerns surrounding Medicare advantage upcoding. http://www.fiercehealthpayer.com/antifraud/story/florida-fraud-case-highlights-concerns-surrounding-medicare-advantage-upcod/2015-02-17 (2015)

Tomar, D., Agarwal, S.: A survey on data mining approaches for healthcare. Int. J. Bio-Sci. Bio-Technol. **5**(5), 241–266 (2013)

Travaille, P., Müller, R.M., Thornton, D., van Hillegersberg, J.: Electronic fraud detection in the US Medicaid healthcare program: lessons learned from other industries. In: 17th Americas Conference on Information Systems, AMCIS (2011)

Witten, I .H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington (2005)

Yamanishi, K., Takeuchi, J.-I., Williams, G., Milne, P.: On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Min. Knowl. Disc. **8**(3), 275–300 (2004)