Vol. 19, Issue 1

Healthcare Fraud Data Mining Methods: A Look Back and Look Ahead

By Nishamathi Kumaraswamy, MS; Mia K. Markey, PhD; Tahir Ekin, PhD; Jamie C. Barner, PhD, FAACP, FAPhA; and Karen Rascati, PhD

Abstract

Healthcare fraud is an expensive, white-collar crime in the United States, and it is not a victimless crime. Costs associated with fraud are passed on to the population in the form of increased premiums or serious harm to beneficiaries. There is an intense need for digital healthcare fraud detection systems to evolve in combating this societal threat. Due to the complex, heterogenic data systems and varied health models across the US, implementing digital advancements in healthcare is difficult. The end goal of healthcare fraud detection is to provide leads to the investigators that can then be inspected more closely with the possibility of recoupments, recoveries, or referrals to the appropriate authorities or agencies. In this article, healthcare fraud detection systems and methods found in the literature are described and summarized. A tabulated list of peer-reviewed articles in this research domain listing the main objectives, conclusions, and data characteristics is provided. The potential gaps identified in the implementation of such systems to real-world healthcare data will be discussed. The authors propose several research topics to fill these gaps for future researchers in this domain.

Keywords: Medicaid, fraud detection, class imbalance, machine learning, health insurance claims

Healthcare Fraud Introduction

Background and Significance

Caring for health has become more expensive, making both private and public administrators more cost conscious in recent years. Therefore, health decision-makers are actively looking for ways to reduce costs. One such avenue of saving potentially billions of dollars is to avoid and detect healthcare fraud. The National Health Care Anti-Fraud Association[1] conservatively estimates that about 3 percent of our healthcare spending is lost to fraud ($300 billion approximately) yearly. Fraud is a complex and difficult problem. It is important to acknowledge that fraud schemes constantly evolve, and fraudsters adapt their methods accordingly. The earliest account[2] of “fraud” in the healthcare literature is from the 1860s when railway collisions were a frequent occurrence, leading to a controversial condition called “railway spine,” which later became a leading cause of personal injury compensation in rail accidents. These accidental events were made profitable by means of insurance settlements in-court or out-of-court by opportunistic claimants, and these events laid the groundwork for fraud definitions and fraud management in the insurance industry.

Healthcare fraud has evolved in the 21st century and has a varied set of profiles ranging from simple fraud schemes to complex networks. The twin objectives of fraud management have always been

fraud prevention and fraud detection[3] (see the definitions section below). The consequence of submitting a fraudulent claim remains the same: the fraudster is prosecuted by means of sanctions and prosecutions in a court of law. However, the methods used in both prevention and detection have evolved since the 1800s, and so have the methods of detecting fraudulent claimants. With the advances in computing, and the more rapid availability of aggregated datasets in the healthcare domain, there are several opportunities for potential advancements in healthcare fraud management. Despite these advancements, it is very difficult to quantify the number of undetected fraudulent cases that do not get prosecuted. The identified limitations[4] in achieving these advancements are manifold, including using legacy systems in claims processing; processing systems that are siloed due to involvement of multiple entities (e.g., enrollment, approvals, authorizations, claims adjudications); having sensitivity related to healthcare data privacy (e.g., sensitive healthcare domains such as family planning and mental health); and difficulty in proving intent of fraud in litigation settings.

The objectives of this review article are to summarize the methods and approaches used in healthcare fraud detection and to discuss the implementation gaps between the academic literature and real-world use by industry settings. Fraud detection in the literature encompasses data mining (rule-based to advanced statistical methods), over-sampling, and extrapolation techniques. The literature concerning overpayment and sampling estimation are important steps in fraud detection's business workflow and are addressed by Ekin et al. (2018).[5]

Definitions

There are many definitions in the literature and social media regarding what constitutes a healthcare fraud incident. Healthcare fraud is defined as an individual, a group of people, or a company knowingly misrepresenting or misstating something about the type, scope, or nature of the medical service provided, which, in turn, results in unauthorized[6,7] payments.

There is a vast amount of literature[8,9] available on fraud management techniques and models in different industries, such as healthcare, telecommunications, credit card services, insurance, and finance. Fraud management,[10] in theory, is divided into two goals: fraud prevention and fraud detection. Fraud prevention in healthcare can be defined as any action or policy that is in place to prevent any system abuse. For example, there is a Medicaid policy in the state of Texas[11] for outpatient mental health services where certain types of providers, such as psychologists and licensed professional counselors, are limited to billing a combined maximum of 12 hours per day, regardless of the number of patients seen. This policy requirement is in effect to prevent fraud (by means of overbilling in this case) before it occurs. Fraud detection, on other hand, is defined as identifying fraud as quickly as possible once a fraudulent scheme has already been perpetrated.

Fraud Actors, Types, and Facts

Healthcare fraud takes many forms. Some of the more prevalent forms[12,13] are traditional fraud schemes implemented by shell vendors, ghost employees who obtained access to bill payers, and employees who continue billing with expired licensures. Some of the main actors committing or involved in fraud are providers (those who are authorized to provide services to beneficiaries), beneficiaries (those who receive medical or associated services), medical equipment manufacturers, drug manufacturers, and agencies authorized to provide special services, such as home healthcare.

Some of the healthcare fraud schemes commonly discussed in literature and used often to develop fraud detection algorithms or analytics within regulatory entities such as the Office of Inspector General (OIG), the Department of Justice (DOJ), and the Centers for Medicare and Medicaid Services (CMS) are as follows:

- Diagnosis Related Groups (DRG) creep – when actors manipulate diagnostic and procedural codes to increase reimbursement amounts in an institutional setting
- Unbundling and fragmentation of procedures – billing individual service codes versus group service codes
- Up-coding of services – billing for a higher level of service than provided
- Phantom billing – billing for services not rendered to clients
- Excess number of services – billing unnecessary services that could lead to client harm
- Kickback schemes – actors might improperly pay for or waive the client's out-of-pocket expense to make up for that cost in additional business
- Billing for mutually exclusive procedures
- Duplicate claims
- Billing errors

Figure 1 illustrates the percentages of improper payments in the United States Health & Human Services (HHS) government programs from 2012 to 2019. Such improper payments include any kind of underpayment, overpayment, fraud, and any unknown payments. The government healthcare programs that were included from the original data source[14] are the following HHS agency programs: Children's Health Insurance Program (CHIP), Medicaid, Medicare Fee-For-Service (FFS), Medicare Part C, and Medicare Part D. As seen in **Figure 1**, the Medicaid and CHIP programs have generally shown a steady increase in the percentage of improper payments.

Figure 2 reports the recoveries from the False Claims Act[15] in years 1985 to 2020. In 2020 alone, $2.2 billion was recovered by the government, out of which $1.8 billion was from the healthcare industry. The recoveries are estimated to be significantly higher for 2021-2022 considering the ongoing difficulties in litigations in closed-court settings due to COVID-19.

Scope and Objectives

The scope of this article is twofold: to provide a comprehensive review of current healthcare-related fraud detection methods and to provide a discussion on implementation gaps in the application of such methods to real-world settings in the US. Related work section entails a comparative evaluation of review studies in literature. This is followed by a review of study methods section, which details selected fraud detection methods with discussions around gaps in applying these methods to real-world data. The next section focuses on implementation gaps, followed by conclusions and future research section, which summarizes the main points and future research directions for healthcare fraud detection. **Table 1** includes an extensive (not exhaustive) tabulated summary of healthcare fraud literature for prospective researchers in this area.

The literature reviewed here does not incorporate articles that included holistic healthcare as an objective, such as those of disease prediction, readmission, or length of stay, in which fraud identification is not necessarily the primary objective. In addition, only articles pertaining to healthcare fraud in the US were considered. In contrast to prior review articles,[16-19] this article discusses the literature from a business workflow perspective starting from a data-driven lead to the end point of litigation/recoupment, and provides recommendations to address the research gaps in existent methods.

Related Work

The value of this review is not only for comparative purposes on the methods employed in the literature but, more importantly, to start a discussion of how relevant current academic healthcare fraud detection methods are to the downstream process of proving intent of fraud by investigators in

an industry setting. An understanding of the implementation gaps and overall fraud detection process (i.e., starting from data leads provided by a model to a conviction phase in a legal setting) will help leverage the already available collective knowledge to help improve practical fraud detection methods.

Several articles discussed healthcare fraud data-mining methods in the literature with similar goals but from different perspectives. Li et al. (2008)[20] categorized the three different actors in healthcare fraud—namely, providers, patients, and the payers—and focused on the provider fraud literature. They further highlighted the scarcity in the data pre-processing methods (from raw claims datasets to flattened datasets) and commented on the importance of this step in identifying healthcare fraud using supervised and unsupervised methods. They also highlighted the two main types of classifier performance metric categories; 1) the error-based methods and 2) the cost-based methods, with error-based classifiers being more common in healthcare fraud literature. An article by Bauder et al. (2017)[21] focused specifically on up-coding fraud in several healthcare domains using medical claims data. They highlighted the lack of literature pertaining specifically to using supervised techniques in up-coding fraud detection.

Ekin et al. (2018)[22] provided a comprehensive discussion of statistical methods in healthcare fraud, including sampling, over-payment estimation methods, and data-mining methods such as supervised, unsupervised, and outlier detection methods from the literature. The authors focused on describing unsupervised methods in more detail, such as using concentration functions and Bayesian co-clustering. Both Ekin et al. (2018)[23] and Li et al. (2008)[24] highlighted the lack of literature in identifying the potential drivers of fraud.

The most recent review by Ai et al. (2021)[25] discussed medical fraud detection methods in the literature using qualitative methods. They provided a methodological literature search using Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines on the methods, number of peer-reviewed articles and a qualitative analysis of statistical methods, model performance, using evaluation metrics (when available) for health care domain. Their research is quite comprehensive, with a focus on being able to assess the strength of model performance and accuracy of existent fraud detection methods in the literature. They concluded that the evidence to provide a consolidated best method to identify healthcare fraud was inadequate considering the literature models were applicable to different domains within healthcare and therefore not directly comparable. They also highlighted that there was no literature available to estimate the cost of investigations in order to estimate potential cost savings using a fraud detection model.

Healthcare administration and payments have changed in the last two decades, especially from a data quality and data integration perspectives. Although the standard forms, such as the CMS-1500 or the UB-04 used for data collection (for payment processing), have not changed significantly over time—except for the volume increase in electronic submissions in the past two decades—there is still a significant gap in the application of literature methods to real-world settings. Other published review articles in this domain focused on the overall state of healthcare fraud literature and methods. This review extends the available literature by focusing on the applicability of these methods to real-world claims data and highlights the research gaps in the practical implementation of these methods.

Policy Statutes Overview

A range of civil, criminal penalties and laws exist within healthcare fraud.[26,27] Government agencies such as the US Department of Justice (DOJ) and the HHS Office of Inspector General (OIG) are the enforcers of such laws and penalties. A quick overview of these laws would aid the understanding of the end goals of the fraud detection business workflow in real-world cases.

The business workflow starts from converting data-based fraud leads to a civil or criminal case indictment, depending on the path an investigation takes, followed by legal proceedings on a case-by-case basis. Data-driven fraud detection tools are only a piece of the complete fraud puzzle; nevertheless, it is an important part considering this is a targeted methodological means to find fraud leads. A simplistic business workflow of how a fraudulent case progresses through a normal course of an investigation/audit is shown in **Figure 3**. The pictograph identifies the most relevant and helpful analytical methods used to identify fraud, waste, or abuse among provider or client or payer.

The common statutes under which fraudulent cases are prosecuted include both civil statutes (the False Claims Act and the Physician Self-Referral Law) and criminal statutes (Anti-Kickback Statute and Criminal Healthcare Fraud Statute).

False Claims Act[28,29] – Many of the fraud cases are lawsuits filed under the False Claims Act (FCA). This is a federal statute originally enacted in the 1800s, and penalties could include recovery of up to three times the damages sustained by the government, in addition to financial penalties for each falsely submitted claim. Most fraudulent recoupments reported by DOJ are claimed under this act.[30]

Physician Self-Referral Law or Stark Law[31] – Under this law, a physician is prohibited from referring patients to receive "designated health services" to an entity in which the physician or immediate family member of the physician has an investment.

Anti-Kickback Statute[32] – Under this law, a medical provider is prohibited from soliciting or receiving any remuneration or rewards directly or indirectly for patient referrals or business generation from anyone.

Criminal Healthcare Fraud Statute[33] – Under this law, any service provider is prohibited from executing a scheme in connection with delivery of health care benefits or services to defraud a health care program.

Data Sources

Healthcare data, in general, are broadly categorized as practitioners' data, administrative claims data, and clinical data.[34] The three sources of data together form a near-complete picture of the fraud data puzzle. However, it is extremely difficult to be in possession of all three data sources under one entity. Second, even if data are available from all three sources, integration of these sources of data can be extremely challenging in real-world practice due to the varied systems and identifiers involved in the data collection and ETL (extract, transform, and load) process. For purposes of fraud detection, the most commonly used data source in the literature is administrative claims.

The collected administrative claims data among insurers do not differ much in their basic structure because of the standard template used in the electronic claims processing. For example, the CMS 1500 form is used in the adjudication process of all professional claims. However, not all collected data are utilized for purposes of adjudication; hence, some data/field values can be considered informational. The data collection and utilization of such informational column values are also dependent on the payer (e.g., fee-for-service versus managed care organization in different state and federal programs). In the next section, the current state-of-the-art fraud detection and prevention methods is briefly described.

Most fraud detection/prevention models discussed in literature are based on either synthetic data or data collected in a de-identified manner and made available as open-source or agency-specific data, such as Veterans Affairs TRICARE, Health and Human Services, or Texas Department of State Health Services. For example, aggregated Medicare/Medicaid data are now made available through

the CMS.gov[35] website. The Medicaid Analytic eXtract contains data collected by CMS from all states on a quarterly basis. Such data are available for researchers to study utilization patterns such as healthcare resource utilization or disease-based utilization. The fraud detection models developed using such aggregated data extracts are difficult for relevant parties to adopt due to the many logistical issues involved, such as the difficulty in linking results tied to the identified provider back to specific claim-line level data.

Rule-Based Fraud Detection

One of the most common approaches to identify fraud is to use domain or expert knowledge to identify anomalies in billing practices. Expert knowledge is often used and is very effective in keeping common fraud schemes in check.

Some common healthcare fraudulent claims as seen in literature fall into the categories mentioned earlier. Simple to medium-complex rules are developed to identify billing errors or duplicate claims to identify fraud categories such DRG creep or up-coding.[36] These are not to be confused with edits and audits in a claims processing system, as these rules are developed based on schemes rather than policy. These rules can be developed at a transaction level or actor level. This is a straightforward and effective approach even though static in nature.

The inherent limitation with such rule-based detection is that once the fraudster becomes aware of the rules—either due to unpaid/rejected/held out claims, or due to a retrospective inspection or audit of adjudicated claims—their fraudulent patterns could change, and these rule-based detection programs cannot quickly adapt to the fraud pattern modifications. Other limitations to having a rule-based detection system are that these engines are very expensive to build, as they require constant inputs from fraud experts and are quite difficult to maintain and manage in the fast-changing healthcare landscape. It is thus very difficult to keep a rule-based system lean and up to date.

Data-Driven Fraud Detection

Data-driven fraud detection is becoming commonly popular in all domains, and the healthcare domain is no exception. Implementing data-driven fraud detection methods offers a higher fraud detection power along with operational and cost efficiencies. The fraud literature regarding the applications of advanced statistical techniques in various healthcare domains (medical, dental etc.) covers three main aspects of the business process: fraud detection, statistical sampling, and oversampling estimation methods. Fraud detection methods[37-42] all have one common motivation, which is to mine data to assess patterns.

Data-driven methods can be categorized broadly as supervised, unsupervised, and hybrid learning methods. These techniques can be summarized from a fraud perspective as below:

- Supervised learning methods employ samples of previously known fraudulent and legitimate transactions or providers.
- Unsupervised learning methods do not require a prior knowledge of fraudulent transactions or providers. They focus more on anomalies based on distributions of a provider's billing behavior. They also use descriptive statistics to help learn such patterns in some cases.
- Hybrid learning is where a mix of both supervised and unsupervised techniques are used.

It is also worth mentioning that these data-mining methods are dependent on a well-defined problem statement and the acquisition of relevant, adequate, and clean data. The process flow of modeling (irrespective of the learning methods used) involves a sequence of steps as it relates to fraud and is

described in **Figure 4**. The different level of complexities involved in a data-driven fraud models from literature are discussed in the next section.

Review of Study Methods in Healthcare Fraud

This section presents selected study methods and discusses practical implementation gaps of these methods. The studies were screened from a structured database search using search terms such as “fraud,” “healthcare,” “secondary data,” “prescriptions,” “Medicaid management information system,” “Medicaid,” “Medicare,” and any possible combinations of these search terms. From this, the studies were further narrowed down focusing on the data, methods, and implementation of fraud algorithms. A subset of such studies are discussed in this section, as they attempt to address some implementation gaps such as class imbalance in real-world data, missing fraud labels, and data pre-processing techniques before applying algorithmic models to data.

Supervised Learning

A supervised learning task is to learn a function that maps response variables to the inputs based on the available labeled response data. Researchers using supervised learning methods in fraud detection have the following in common: a labeled dataset (i.e., fraudulent: yes or no), a domain-specific justification to choose one algorithm versus another, and a performance metric of choice to determine the best algorithm. The general concept that stands out in the development of such supervised models is the identification of features that can discriminate a fraudulent provider from legitimate providers. The methods of identifying such features vary between researchers and are mostly focused from a provider-level rather than a transaction-level.

Considerations in Defining Ground Truth

It is important to acknowledge that any supervised technique application is inherently dependent on the validity of the labeled dataset used to categorize the data to their corresponding classes. Supervised learning algorithms thus require confidence in the correct classification/labeling of the providers. The fraud labels for the reviewed providers are classified to one of two categories: fraudulent or not fraudulent (legitimate). But it is not known if providers who were never reviewed did or did not commit fraud. Some published studies[43-45] address this uncertainty partially by having a varied range as an estimate for class distribution of the “never reviewed” providers. Thus, there will always be cases where fraud is mislabeled as non-fraud. Binary classification of providers as fraudulent or legitimate does not allow for uncertainty to remain after providers are investigated. In contrast, the confidence that a provider committed fraud (“fraud” confidence) could be used for supervised learning in lieu of a binary ground truth.

The labeled fraud dataset is skewed in nature, irrespective of methods used for label associations in a dataset. The skewness arises from the practical fact that only a small number of the reviewed providers are categorized as fraudulent while the majority of the reviewed providers are legitimate. This nature of skewness in a categorical label assignment is called “class imbalance” and has its own literature[46] stemming from computer science and its applications to real-world problems.

Review of Supervised Learning in Healthcare Fraud Detection

Bauder et al. (2018 and 2018, May)[47,48] categorized different supervised learning techniques (Random Forest, C4.5 decision tree, support vector machine, and logistic regression) to find the effect of class imbalance in fraud detection. The authors used publicly available claims data (Medicare Provider Utilization and Payment Data: Physician and Other Supplier) from CMS. The

labels for known fraudulent medical providers across all specialties, and provider types were obtained from the OIG's publicly available database of List of Excluded Individuals/Entities (LEIE) in 2017. The final merged Medicare dataset (claims and labeled fraud data) was highly imbalanced (about nine out of every 100,000 providers were marked fraudulent). The performance metrics used were area under curve (AUC); false positive rate (FPR is the ratio of non-fraud cases incorrectly categorized as fraudulent cases to the total number of non-fraudulent cases); and false negative rate (FNR is the ratio of fraud cases incorrectly categorized as non-fraud cases to the total number of fraudulent cases). Two main conclusions were:

- The C4.5 (decision tree) algorithm had the best performance on the AUC metric (0.883).
- As the minority class distribution was varied from 20 percent to 50 percent, the learners became worse on their performance metrics.

Herland et al. (2018 and 2019)[49,50] also investigated the effects of class imbalance on supervised learning for fraud detection using the same publicly available datasets (claims and fraud labels) as Bauder et al. (2018, May). The authors concluded that a logistic regression model followed by gradient tree boosting performed well based on the AUC metric (0.828) evaluation.

Fan et al. (2019)[51] focused on physician fraud detection combining the different open datasets on claims (CMS data), social media ratings on physicians (Healthgrades.com), and ground truth fraud datasets such as LEIE and Board Actions. The different classifiers that were trained included logistic regression, naïve Bayes, and a decision tree classifier. The board action dataset features did not prove to be beneficial to their classification model, although it is not clear which features from the dataset were included in the modeling process. Some feature engineering was performed to determine the final set of features resulting in a best classifier. The authors concluded that their classification performance was highest using a decision tree with features (based on rating) from social media, open payment, and prescriber (CMS) datasets.

Ekin et al. (2021)[52] provided an overview of pros and cons in addressing three steps of the statistical fraud detection modeling process. In their experimental design, they manipulated the claims data to address the variance in the model performance from:

1. Correlated features – e.g., principal component analysis (PCA) on the features to address multicollinearity

2. Classifier type – nine supervised classification algorithms such as random forest, naïve Bayes, and neural networks.

3. Class imbalance – this effect was addressed by using four sampling techniques (e.g., random walk oversampling (RWO))

They utilized a wide range of evaluation metrics to assess the different model's performance with the aggregated public (CMS's Part B, CMS's zipcode to carrier locality file, and CMS's Geographic Variation Public Use File) datasets. To simulate an adjustment to the well-known method of considering LEIE data as the only source of ground truth for fraud labels, they performed an experiment with a range of possible fraud proportions (0.06 percent to 45.76 percent). The combination of these data manipulations led to a total of 405 different trained models. Based on their AUC metric (0.84) performance, their conclusions were:

1. As class imbalance becomes higher, AUC becomes lower. [This is in contrast to the first three articles that were discussed above and more in agreement with the imbalanced data literature].

2. The best sampling approach with the highest AUC was RWO. [This has not been previously reported in literature by any other published studies and is an important addition to the fraud literature].

3. Overall correlated features did not affect model performance for most of the algorithms; however, the authors do report a slight improvement in performance using PCA for random forest algorithms. [This is slightly inconsistent to the familiar belief in data science that correlated features affect model stability for some algorithms and will need a larger training data set].[53]

One common theme among these methods is that these authors used publicly available datasets and LEIE or expert opinions as their ground truth. However, the ground truth was considered binary for all these studies. Another major limitation in these studies was that the features used to train the models did not extend beyond what is available in the aggregated data extract, which limits feature engineering. Lastly, the evaluation metrics used were all error-based rather than cost-based. However, no prior research exists to make an estimate on cost and resources needed for an investigation.

Unsupervised Learning

Unsupervised learning refers to techniques that are used to identify patterns or structure in data. These descriptive techniques are used when no labels or class markers are available for the algorithm to learn from. The K-means clustering technique is a commonly seen example of unsupervised learning. These methods can also be a precursor to use before descriptive outlier techniques can be implemented. A more detailed overview of such methods can be found in Konjin (2017).[54]

Ekin et al. (2013)[55] proposed a novel Bayesian co-clustering framework to detect healthcare fraud. In this theoretical framework, the authors identify a co-cluster (defined as a dyadic grouping of provider and beneficiary) as fraudulent depending on the posterior probability assigned to the co-cluster. A Dirichlet distribution prior and Beta distribution prior for the random variable were proposed to arrive at the posterior probability distribution of each co-cluster. Any co-cluster containing unusual membership of either provider or a beneficiary was identified as fraudulent. The authors only used simulated data to test the feasibility of this framework. Nevertheless, this is an important step in the direction of investigating conspirator fraud involving two parties.

Sadiq et al. (2017)[56] used CMS Medicare Part B, Part D and durable medical equipment, prosthetics, orthotics (DMEPOS) datasets to develop a fraud claim detection system using the Patient Rule Induction Method (PRIM) based bump hunting method. PRIM starts with all of the training data and peels/removes regions followed by paste/addition of regions, thereby gradually zooming into regions with high values for target variables. At each step of the peel or paste, only a small set of observations are excluded or included in this heuristic search. For example, when a physician prescribed eye drops, they calculated the conditional probability of the prescriber being an ophthalmologist. A low probability indicated a higher likelihood that the prescription was improper. The homogeneity of prescriptions (overused medications for conditions such as headache, cold) was accounted for by using an analysis of variance (ANOVA) test and F score calculations. The identified bump regions were validated by comparing the number of observations in bump region with that not in the bump region using a confusion matrix. It is unclear from the article how the ground truth for the bump region was determined. This method may help in identifying potential fraudsters who may go unidentified when using other popular classifiers.

Sadiq et al. (2019)[57] used propensity matching and clustering for fraud detection using CMS's 2012-2015 Medicare Part B, Part D and DMEPOS dataset. They term this temporal learning framework as Cascaded Propensity Matching (CPM) Fraud Miner. Their primary goal was to see if a deliberate fraudulent action causes a perturbation in the observational data, accounting for any co-variates (X) that could lead to that fraudulent outcome. A weighted propensity score e(X) was used to compare the treatment group (T=1, indicating patients were treated by other physicians for a given condition) with the other-treatment group (T=0, indicating patients were not treated by other physicians for a given condition). For example, a condition such as a cataract is never treated by retinoblastoma removal surgery, but there could be situations or clients where such a condition is treated by this expensive surgery. The reason for such an expensive option billed in the data could be twofold: due to sheer neglect of the condition by the clients for many years, including other medical conditions leading to have the surgery, or due to a high reimbursement amount to providers for expensive surgeries. However, the method used to arrive at the ground truth or how these database results were incorporated to determine their performance metrics/values was unclear.

Zafari and Ekin (2019)[58] devised an unsupervised framework for prescription fraud using 2015 Part D Medicare data. The prescribing specialty code was used as a co-variate to control for the deviations normally evident in prescriptions due to the prescriber's specialty. These authors proposed a novel approach to detect associations between prescribers and drug group (topics) billings from transactional data, accounting for specialty differences that could lead to different billings. They used structural topic modeling concepts (from natural language processing literature) to group drugs into different topics (synonymous to grouped drug categories) for all prescribers. These groupings were then used as benchmark groups to detect outliers by means of concentration functions or distance-based measures that capture deviations from expected billing patterns. Their fraud framework can be summarized as consisting of two main steps: Step 1 involves identifying associations between prescribers and their drug billing utilization, followed by Step 2, which uses these groupings to detect outliers within a specialty.

Ekin et al. (2019)[59] outlined a novel unsupervised Bayesian hierarchical model to help untangle the relationship between medical procedures and billing provider using CMS's Part-B data. The joint distributions of these variables were derived, and a Gibbs sampler was used to draw samples from the posterior density function of this joint distribution. Their unsupervised hierarchical Bayesian framework implementation using real-world data identified anomalies in billing among providers who could be outliers considering only the billing procedure code and frequency of billing by the provider. One possible limitation that might explain some of the anomalies could be the lack of consideration for billing modifiers (modifiers allow a provider to indicate special circumstantial usage of a performed procedure) in a claim, but Part-B data does not contain modifiers, which is an inherent limitation of the data source.

Such outlier detection techniques allow for real-world implementation without having to rely on a labeled fraud dataset; however, the burden of proving intent of fraud lies with the investigators or experts in these cases. Identification of the claim line details that are the root causes for such provider billing anomalies is difficult considering that aberrant behavior of billing exists for the provider as a whole. These methods used, along with rule-based outlier detection techniques, could overcome some limitations of unsupervised techniques. Another major limitation of these methods in the real world is the governing business workflow for a fraud examiner.

Hybrid Learning

Hybrid learning includes a mix of both supervised and unsupervised techniques that are used at different stages of the model.

Shin et al. (2012)[60] proposed a tree-based segmentation model and scoring using outlier techniques and a method that unifies these different techniques to provide a final score. Eldardiry et al. (2013)[61] used rules to identify features, followed by rule-based feature extraction and risk score computation using the term frequency-inverse document frequency method that reflected how important or unimportant a particular rule violation is for a provider. In summary, Eldardiry et al.[62] applied a probabilistic outlier detection technique by combining features from a set of fraud rules to design a risk score computation, which would allow an investigator to tease out the specific rule a provider violated in the course of an investigation.

Identified Recommendations and Implementation Gaps

The literature justifies the applicability of machine learning and statistical techniques to healthcare provider fraud detection. Most research describes the process of utilizing open source aggregated transactional health data merged with some kind of ground truth (e.g., OIG exclusion list, expert opinions) database and validates their model using known performance metrics such as F-score, or recall. Some authors focus on specific areas such as prescription medications or DMEPOS as their data source. Different authors use different techniques ranging from a rule-based technique to advanced statistical models using algorithms such as multinomial naïve Bayes or logistic regression.

In spite of the extensive literature in this area, there is no unifying process that bundles these research methods together to identify healthcare fraud. There is, however, a need to have a unified framework to provide a solution to the problem of retrospective fraud detection in healthcare domains such as prescription claims, dental claims, long-term care service claims, and professional claims. This process should facilitate easy integration into an investigator or auditor's workflow so as to meet the business and real-world needs of investigations.

Second, feature engineering is highly dependent on the data sources. This dependency on data sources inherently poses limitations on the number of co-variates considered as being associated with our response variable (provider fraud) in different domains. For example, provider specialty is a covariate that is usually considered in professional claims analysis, while provider specialties in the case of pharmacies such as chain, independent, mail service, and wholesalers are not considered in prescription claims analysis. There is a need for algorithms to be adaptive to the number of features available in different state or private payer systems and data sources within a domain. The model needs to be able to accommodate and degrade gracefully, dependent on the feature space considered for the focused healthcare domain. "Graceful degradation" here is defined as the ability of a model to maintain limited functionality even when a portion of data, or some variables, are missing. Most healthcare fraud literature does not discuss further feature engineering or data pre-processing from raw line-level claim element attributes. This is a common gap seen in this literature. Feature engineering is a key factor known to affect algorithmic performance to a great extent in the computer science literature. This is an inherent limitation of the aggregated data sources often used in fraud literature.

Third, fraud models in the literature seem to start with aggregated data before using a predictive algorithm on the collected or processed data. It is unclear how certain components of claims were dealt with in these aggregations. For example, in prescription claims, the quantity dispensed and whether or not the product was compounded are variables available for each claim (each prescription dispensed), but the method of aggregation of these variables remains unclear. A new train of thought is needed here to explore the idea of extracting patterns or meaning from de-aggregated claims. This will help build into the existing body of fraud literature and help build models that can be used for real-world investigations.

Lastly, the fraud literature does not address the actions or processes that happen beyond fraud-detection. Prospective researchers in this domain have the opportunity to follow the results beyond detection. Qualitative analysis, employing focus groups, and interviews with investigators, auditors, and litigators who play a very important role beyond the stage of fraud identification (using advanced statistical methods) is crucial to improving algorithms that will provide results that can be integrated into the business workflow process.

These gaps in healthcare fraud research are portrayed pictorially in **Figure 5** for one healthcare domain (pharmacy) for demonstration purposes.

Conclusions and Future Research Directions

Even though the academic literature on statistical methods used to identify healthcare fraud is substantial, there are very few states in US that implement these advanced methods in real-world practice. Possible reasons for the lack of application are discussed below:

1. The structure of such teams focused on fraud detection from a business workflow standpoint consists of a team of trained and credentialed auditors, administrative/criminal investigators, statisticians/analysts/both, and investigative attorneys within any state or federal integrity programs. Considering this business workflow (see **Figure 3**), there is a strong need for collaboration of the data team (statisticians/analysts/both) and the examiner's (auditors, investigators, and attorneys) team to identify and convert fraud leads to recoupments (fraud conversion rate).
2. There is also a strong need for closing the feedback loop on what worked and what did not from an investigation and litigation standpoint. This information collected in a quantitative or qualitative fashion (e.g., focus groups, interviews) can help fine-tune mining algorithms leading to an improved fraud conversion rate.

3. There is also a significant gap in implementing the methods reviewed in this article in real-world use cases since fraudulent intent is difficult to prove, and without fraudulent intent, the actors cannot be prosecuted.

4. Complex algorithms are difficult for the downstream examiner's team to understand and use. In a healthcare fraud business workflow, it is very important that the methods used in each step along the way are transparent and easy to comprehend. Such logistical issues are hypothesized in impeding progress from algorithm design to implementation.

5. Most methods in the literature use publicly available data, which is a major limitation to implementation in a business setting. However, other methods that have used private data sources, such as those from electronic health records or private payer data, are limited due to data privacy and legal issues and are thus difficult to replicate to a real-world setting.

6. Drilling down from provider-level data-based leads (using advanced statistical methods) to specific claim-level leads is important for an investigator to make a case of fraud. This gap might be addressed by educating investigators on pursuing such data-based provider leads from an investigation standpoint in collaboration with the data team.

The ability of a method to tease out the metrics that best identify a fraudulent provider lead needs some fine-tuning as well. The literature included different modeling metrics to assess a model; however, from a practical implementation standpoint, research on the costs associated with a fraud investigation is not quantified. This is a significant gap that needs to be addressed so the statistical methods to identify fraud can be modified from an error-based to a cost-based solution.

In conclusion, this article reviewed healthcare fraud detection systems and methods found in the academic literature, discussed limitations and implementation gaps of such methods to real-world business setting, and concluded with an outline of potential solutions to address these gaps.

Support

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Notes

1. National Health Care Anti-Fraud Association, “The Challenge of Health Care Fraud.” 2021. **https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/**

2. Dornstein, Ken. “Accidentally, on purpose: The making of a personal injury underworld in America.” St. Martin’s Press, 1996.

3. Bolton, Richard J., and David J. Hand. “Statistical fraud detection: A review.” *Statistical Science* 17, no. 3 (2002): 235-255.

4. Travaille, Peter. “Electronic fraud detection in the US Medicaid Health Care Program.” Master’s thesis, University of Twente, 2011.

5. Ekin, Tahir, Francesca Ieva, Fabrizio Ruggeri, and Refik Soyer. “Statistical medical fraud assessment: exposition to an emerging field.” *International Statistical Review* 86, no. 3 (2018): 379-402.

6. CFO (2018). United States Chief Financial Officer’s Council - Anti Fraud Playbook. **https://www.cfo.gov/knowledge-sharing/fraudprevention/**. Accessed on July 26, 2021.

7. USAO (2019). United State Attorney’s Office Western District of Michigan. **https://www.justice.gov/usao-wdmi/health-care-fraud**. Accessed on July 26, 2021

8. Ekin, 2018.

9. Li, Jing, Kuei-Ying Huang, Jionghua Jin, and Jianjun Shi. “A survey on statistical methods for health care fraud detection.” *Health Care Management Science* 11, no. 3 (2008): 275-287.

10. Bolton, 2002.

11. TMHP (2021). Texas State Medicaid - Behavioral Health and Case Management handbook, Texas Medicaid Provider Procedures Manual: Vol. 2. Section 4.5 **https://www.tmhp.com/sites/default/files/file-library/resources/provider-manuals/tmppm/pdf-chapters/2021/2021-07-july/2_Behavioral_Health.pdf**. Accessed on July 26, 2021.

12. Ahadiat, Nas, and Mohamed Gomaa. “Healthcare Fraud and Abuse: An Investigation of the Nature and Most Common Schemes.” *Journal of Forensic and Investigative Accounting* 10, no. 3 (2018): 428-435.

13. Stowell, Nicole F., Martina Schmidt, and Nathan Wadlinger. “Healthcare fraud under the microscope: improving its prevention.” *Journal of Financial Crime* (2018).

14. Payment Accuracy (2020). Dataset downloads, Annual Improper Payments Datasets - Payment Accuracy 2020 Dataset. **https://www.paymentaccuracy.gov/payment-accuracy-the-numbers/**. Downloaded on July 26, 2021.

15. DOJ (2021). Datasets downloaded as PDF. **https://www.justice.gov/opa/pr/justice-department-recovers-over-22-billion-false-claims-act-cases-fiscal-year-2020**. Accessed: July 26, 2021.

16. Ekin, 2018.

17. Li, Jing, 2008.

18. Bauder, Richard, Taghi M. Khoshgoftaar, and Naeem Seliya. “A survey on the state of healthcare upcoding fraud analysis and detection.” *Health Services and Outcomes Research Methodology* 17, no. 1 (2017): 31-55.

19. Ai, Jing, Jennifer Russomanno, Skyla Guigou, and Rachel Allan. “A Systematic Review and Qualitative Assessment of Fraud Detection Methodologies in Health Care.” *North American Actuarial Journal* (2021): 1-26.

20. Li, Jing, 2008.

21. Bauder, Richard, Taghi M. Khoshgoftaar, and Naeem Seliya. “A survey on the state of healthcare upcoding fraud analysis and detection.” *Health Services and Outcomes Research Methodology* 17, no. 1 (2017): 31-55.

22. Ekin, 2018.

23. Ibid.

24. Li, Jing, 2008.

25. Ai, Jing, 2021.

26. Fabrikant, Robert, Paul E. Kalb, Pamela H. Bucy, and Mark D. Hopson. Health care fraud: Enforcement and compliance. *Law Journal Press*, 2021.

27. Pacini, Carl, Nicole Forbes Stowell, and Maria T. Caban-Garcia. "A Forensic Accountant’s Guide to the Most Potent Federal Laws Used Against Healthcare Fraud." Pacini, CJ, Sowell Forbes, N. and Cabán-García, MT (2020): 386-406.

28. Ibid.

29. Salcido, Robert. “The government’s increasing use of the False Claims Act against the health care industry.” *The Journal of Legal Medicine* 24, no. 4 (2003): 457-494.

30. DOJ (2021).

31. Pacini, Carl, 2020.

32. Ibid.

33. Ibid.

34. Liu, Qi, and Miklos Vasarhelyi. “Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information.” 29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia. 2013.

35. CMS RESDAC (2021). Datasets available through Research Data Assistance Center. **https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAXGeneralInformation**. Accessed on July 26, 2021

36. Bauder, Richard, Taghi M. Khoshgoftaar, and Naeem Seliya. “A survey on the state of healthcare upcoding fraud analysis and detection.” *Health Services and Outcomes Research Methodology* 17, no. 1 (2017): 31-55.

37. Rosenberg, Marjorie A., Dennis G. Fryback, and David A. Katz. “A statistical model to detect DRG upcoding.” *Health Services and Outcomes Research Methodology* 1, no. 3 (2000): 233-252.

38. Phua, Clifton, Vincent Lee, Kate Smith, and Ross Gayler. “A comprehensive survey of data mining-based fraud detection research.” arXiv preprint arXiv:1009.6119 (2010).

39. Shin, Hyunjung, Hayoung Park, Junwoo Lee, and Won Chul Jhee. “A scoring model to detect abusive billing patterns in health insurance claims.” *Expert Systems with Applications* 39, no. 8 (2012): 7441-7450.

40. Capelleveen, Guido Cornelis. “Outlier based predictors for health insurance fraud detection within US Medicaid.” Master’s thesis, University of Twente, 2013.

41. van Capelleveen, Guido, Mannes Poel, Roland M. Mueller, Dallas Thornton, and Jos van Hillegersberg. “Outlier detection in healthcare fraud: A case study in the Medicaid dental domain.” *International Journal of Accounting Information Systems* 21 (2016): 18-31.

42. Eldardiry, Hoda, Juan Liu, Ying Zhang, and Markus Fromherz. “Fraud detection for healthcare.” Knowledge Discovery and Data Mining Workshop on Data Mining for Healthcare. 2013.

43. Ekin, Tahir, Luca Frigau, and Claudio Conversano. “Health care fraud classifiers in practice.” *Applied Stochastic Models in Business and Industry* (2021).

44. Bauder, Richard, and Taghi Khoshgoftaar. “Medicare fraud detection using random forest with class imbalanced big data.” In 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 80-87. IEEE, 2018.

45. Bauder, Richard A., and Taghi M. Khoshgoftaar. “The detection of medicare fraud using machine learning methods with excluded provider labels.” The Thirty-First International Flairs Conference. 2018.

46. Ali, Haseeb, Mohd Najib Mohd Salleh, Rohmat Saedudin, Kashif Hussain, and Muhammad Faheem Mushtaq. “Imbalance class problems in data mining: A review.” *Indonesian Journal of Electrical Engineering and Computer Science* 14, no. 3 (2019): 1560-1571.

47. Bauder, Richard, and Taghi Khoshgoftaar. “Medicare fraud detection using random forest with class imbalanced big data.” In 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 80-87. IEEE, 2018.

48. Bauder, Richard A., and Taghi M. Khoshgoftaar. “The detection of medicare fraud using machine learning methods with excluded provider labels.” The Thirty-First International Flairs Conference. 2018.

49. Herland, Matthew, Taghi M. Khoshgoftaar, and Richard A. Bauder. “Big data fraud detection using multiple medicare data sources.” *Journal of Big Data* 5, no. 1 (2018): 1-21.

50. Herland, Matthew, Richard A. Bauder, and Taghi M. Khoshgoftaar. “The effects of class rarity on the evaluation of supervised healthcare fraud detection models.” *Journal of Big Data* 6, no. 1 (2019): 1-33

51. Fan, Brandon, Xuan Zhang, and Weiguo Fan. “Identifying physician fraud in healthcare with open data.” In International Conference on Smart Health, pp. 222-235. Springer, Cham, 2019.

52. Ekin, Tahir, 2021.

53. Toloşi, Laura, and Thomas Lengauer. “Classification with correlated features: unreliability of feature ranking and solutions.” *Bioinformatics* 27, no. 14 (2011): 1986-1994.

54. Konijn, R. M. “Detecting interesting differences: Data mining in health insurance data using outlier detection and subgroup discovery.” (2017).

55. Ekin, Tahir, Francesca Leva, Fabrizio Ruggeri, and Refik Soyer. “Application of bayesian methods in detection of healthcare fraud.” *Chemical Engineering Transactions* 33 (2013)

56. Sadiq, Saad, Yudong Tao, Yilin Yan, and Mei-Ling Shyu. “Mining anomalies in medicare big data using patient rule induction method.” In 2017 IEEE Third International Conference on Multimedia Big Data (BigMM), pp. 185-192. IEEE, 2017.

57. Sadiq, Saad, and Mei-Ling Shyu. “Cascaded propensity matched fraud miner: Detecting anomalies in medicare big data.” *Journal of Innovative Technology* 1, no. 1 (2019): 51-61.

58. Zafari, Babak, and Tahir Ekin. “Topic modelling for medical prescription fraud and abuse detection.” *Journal of the Royal Statistical Society: Series C* (Applied Statistics) 68, no. 3 (2019): 751-769.

59. Ekin, Tahir, Greg Lakomski, and Rasim Muzaffer Musal. “An unsupervised Bayesian hierarchical method for medical fraud assessment.” *Statistical Analysis and Data Mining: The ASA Data Science Journal* 12, no. 2 (2019): 116-124.

60. Shin, Hyunjung, 2012.

61. Eldardiry, Hoda, 2013.

62. Ibid.

63. Peng, Yi, Gang Kou, Alan Sabatka, Zhengxin Chen, Deepak Khazanchi, and Yong Shi. “Application of clustering methods to health insurance fraud detection.” In *2006 International Conference on Service Systems and Service Management*, vol. 1, pp. 116-120. IEEE, 2006.

64. Shin, Hyunjung, 2012.

65. Ekin, Tahir, 2013.

66. Eldardiry, Hoda, 2013.

67. Thornton, Dallas, Roland M. Mueller, Paulus Schoutsen, and Jos Van Hillegersberg. “Predicting healthcare fraud in medicaid: a multidimensional data model and analysis techniques for fraud detection.” *Procedia Technology* 9 (2013): 1252-1264.

68. Bowblis, John R., and Christopher S. Brunt. “Medicare skilled nursing facility reimbursement and upcoding.” *Health Economics* 23, no. 7 (2014): 821-840.

69. Joudaki, Hossein, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi, Bijan Geraili, Mahdi Nasiri, and Mohammad Arab. “Using data mining to detect health care fraud and abuse: a review of literature.” *Global Journal of Health Science* 7, no. 1 (2015): 194.

70. Bauder, Richard A., Taghi M. Khoshgoftaar, Aaron Richter, and Matthew Herland. “Predicting medical provider specialties to detect anomalous insurance claims.” In 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 784-790. IEEE, 2016.

71. Joudaki, Hossein, Arash Rashidian, Behrouz Minaei-Bidgoli, Mahmood Mahmoodi, Bijan Geraili, Mahdi Nasiri, and Mohammad Arab. “Improving fraud and abuse detection in general physician claims: a data mining study.” *International Journal of Health Policy and Management* 5, no. 3 (2016): 165.

72. van Capelleveen, Guido, 2016.

73. Bauder, Richard, Taghi M. Khoshgoftaar, and Naeem Seliya. “A survey on the state of healthcare upcoding fraud analysis and detection.” *Health Services and Outcomes Research Methodology* 17, no. 1 (2017): 31-55.

74. Ekin, Tahir, Francesca Ieva, Fabrizio Ruggeri, and Refik Soyer. “On the use of the concentration function in medical fraud assessment.” *The American Statistician* 71, no. 3 (2017): 236-241.

75. Sadiq, Saad, 2017.

76. Bauder, Richard A., and Taghi M. Khoshgoftaar. “The detection of medicare fraud using machine learning methods with excluded provider labels.” The Thirty-First International Flairs Conference. 2018.

77. Bauder, Richard, and Taghi Khoshgoftaar. “Medicare fraud detection using random forest with class imbalanced big data.” In 2018 IEEE International Conference on Information Reuse and Integration (IRI), pp. 80-87. IEEE, 2018.

78. Herland, Matthew, 2018.

79. Fan, Brandon, 2019.

80. Herland, Matthew, 2019.

81. Sadiq, Saad, 2019.

82. Zafari, Babak, 2019.

83. Ekin, Tahir, Luca Frigau, and Claudio Conversano. “Health care fraud classifiers in practice.” *Applied Stochastic Models in Business and Industry* (2021).

84. Ai, Jing, 2021.

85. Wirth, Rüdiger, and Jochen Hipp. “CRISP-DM: Towards a standard process model for data mining.” *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, vol. 1. London, UK: Springer-Verlag, 2000.

86. Azevedo, Ana Isabel Rojão Lourenço, and Manuel Filipe Santos. “KDD, SEMMA and CRISP-DM: a parallel overview.” IADS-DM (2008).

Author Biographies

Nishamathi Kumaraswamy (corresponding author) is a graduate student at the University of Texas at Austin College of Pharmacy.

Mia K. Markey is a professor of biomedical engineering at the University of Texas at Austin in the Department of Biomedical Engineering.

Tahir Ekin is an associate professor of quantitative methods at Texas State University in the Department of Computer Information Systems and Quantitative Methods.

Jamie C. Barner is a professor of health outcomes and pharmacy practice at the University of Texas at Austin College of Pharmacy.

Karen Rascati is a professor of health outcomes and pharmacy Practice at the University of Texas at Austin College of Pharmacy.