Herland *et al. J Big Data (2018) 5:29*
https://doi.org/10.1186/s40537-018-0138-3

Journal of Big Data

RESEARCH Open Access

Big Data fraud detection using multiple medicare data sources

Matthew Herland*, Taghi M. Khoshgoftaar and Richard A. Bauder

*Correspondence:
mherlan1@fau.edu
Florida Atlantic University,
777 Glades Road, Boca Raton,
FL, USA

Abstract

In the United States, advances in technology and medical sciences continue to improve the general well-being of the population. With this continued progress, programs such as Medicare are needed to help manage the high costs associated with quality healthcare. Unfortunately, there are individuals who commit fraud for nefarious reasons and personal gain, limiting Medicare's ability to effectively provide for the healthcare needs of the elderly and other qualifying people. To minimize fraudulent activities, the Centers for Medicare and Medicaid Services (CMS) released a number of "Big Data" datasets for different parts of the Medicare program. In this paper, we focus on the detection of Medicare fraud using the following CMS datasets: (1) Medicare Provider Utilization and Payment Data: Physician and Other Supplier (Part B), (2) Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D), and (3) Medicare Provider Utilization and Payment Data: Referring Durable Medical Equipment, Prosthetics, Orthotics and Supplies (DMEPOS). Additionally, we create a fourth dataset which is a combination of the three primary datasets. We discuss data processing for all four datasets and the mapping of real-world provider fraud labels using the List of Excluded Individuals and Entities (LEIE) from the Office of the Inspector General. Our exploratory analysis on Medicare fraud detection involves building and assessing three learners on each dataset. Based on the Area under the Receiver Operating Characteristic (ROC) Curve performance metric, our results show that the Combined dataset with the Logistic Regression (LR) learner yielded the best overall score at 0.816, closely followed by the Part B dataset with LR at 0.805. Overall, the Combined and Part B datasets produced the best fraud detection performance with no statistical difference between these datasets, over all the learners. Therefore, based on our results and the assumption that there is no way to know within which part of Medicare a physician will commit fraud, we suggest using the Combined dataset for detecting fraudulent behavior when a physician has submitted payments through any or all Medicare parts evaluated in our study.

Keywords: Big Data, U.S. Medicare, LEIE, Fraud detection

Introduction

Healthcare in the United States (U.S.) is important in the lives of many citizens, but unfortunately the high costs of health-related services leave many patients with limited medical care. In response, the U.S. government has established and funded programs, such as Medicare [1], that provide financial assistance for qualifying people to receive needed medical services [2]. There are a number of issues facing healthcare and

medical insurance systems, such as a growing population or bad actors (i.e. fraudulent or potentially fraudulent physicians/providers), which reduces allocated funds for these programs. The United States has experienced significant growth in the elderly population (65 or older), in part due to the improved quality of healthcare, increasing 28% from 2004 to 2015 compared to 6.5% for Americans under 65 [3]. Due, in part, to the increase in population, especially for the elderly demographic, as well as advancements in medical technology, U.S. healthcare spending increased, with an annualized growth rate between 1995 and 2015 of 4.0% (adjusted for inflation) [4]. Presumably, spending will continue to rise, thus increasing the need for an efficient and cost-effective healthcare system. A significant issue facing healthcare is fraud, waste and abuse, where even though there are efforts being made to reduce these [5], they are not significantly reducing the consequent financial strain [6]. In this study, we focus our attention on fraud, and use the word fraud in this paper to include the terms waste and abuse. The Federal Bureau of Investigation (FBI) estimates that fraud accounts for 3–10% of healthcare costs [7], totaling between $19 billion and $65 billion in financial loss per year. Medicare accounts for 20% of all U.S. healthcare spending [8] with a total possible cost recovery (with the potential application of effective fraud detection methods) of $3.8 to $13 billion from Medicare alone. Note that Medicare is a federally subsidized medical insurance, and therefore is not a functioning health insurance market in the same way as private healthcare insurance companies [9]. There are two payment systems available through Medicare: Fee-For-Service and Medicare Advantage. For this study, we focus on data within the Fee-For-Service system of Medicare where the basic claims process consists of a physician (or other healthcare provider) performing one or more procedures and then submitting a claim to Medicare for payment, rather than directly billing the patient. The second payment system, Medicare Advantage, is obtained through a private company contracted with Medicare, where the private company manages the claims and payment processes [10]. Additional information on the Medicare process and Medicare fraud is provided within [1, 11–13].

The detection of fraud within healthcare is primarily found through manual effort by auditors or investigators searching through numerous records to find possibly suspicious or fraudulent behaviors [14]. This manual process, with massive amounts of data to sieve through, can be tedious and very inefficient compared to more automated data mining and machine learning approaches for detecting fraud [15, 16]. The volume of information within healthcare continues to increase due to technological advances allowing for the storage of high-volume information, such as in Electronic Health Records (EHR), enabling the use of "Big Data." As technology advances and its use increases, so does the ability to perform data mining and machine learning on Big Data, which can improve the state of healthcare and medical insurance programs for patients to receive quality medical care. The Centers for Medicare and Medicaid Services (CMS) joined in this effort by releasing "Big Data" Medicare datasets to assist in identifying fraud, waste and abuse within Medicare [17]. CMS released a statement that "those intent on abusing Federal health care programs can cost taxpayers billions of dollars while putting beneficiaries' health and welfare at risk. The impact of these losses and risks magnifies as Medicare continues to serve a growing number of people [18]." There are several datasets available at the Centers for Medicare and Medicaid Services website [8].

In this study, we use three Public Use File (PUF) datasets: (1) Medicare Provider Utilization and Payment Data: Physician and Other Supplier (Part B), (2) Medicare Provider Utilization and Payment Data: Part D Prescriber (Part D), and (3) Medicare Provider Utilization and Payment Data: Referring Durable Medical Equipment, Prosthetics, Orthotics. and Supplies (DMEPOS). We chose these parts of Medicare because they cover a wide range of possible provider claims, the information is presented in similar formats, and they are publicly available. Furthermore, the Part B, Part D, and DMEPOS dataset comprise key components of the Medicare program and by incorporating all three aspects of Medicare for fraud detection, this study provides a comprehensive view of fraud in the Medicare program. Information provided in these datasets includes the average amount paid for these services and other data points related to procedures performed, drugs administered, or supplies issued. We also create a dataset combining all three of these Medicare datasets, which we refer to as the Combined dataset. The last dataset examined in our study is the List of Excluded Individuals and Entities (LEIE) [19], provided by Office of the Inspector General, which contains real-world fraudulent physicians and entities.

The definition of Big Data is not universally agreed upon throughout the literature [20–24], so we use an encompassing definition by Demchenko et al. [25] who define Big Data by five V's: Volume, Velocity, Variety, Veracity and Value. Volume pertains to vast amounts of data, Velocity applies to the high pace at which new data is generated/collected, Variety pertains to the level of complexity of the data (e.g. incorporating data from different sources), Veracity represents the genuineness of the data, and Value implies how good the quality of the data is in reference to the intended results. The datasets released by CMS exhibit many of these Big Data qualities. These datasets qualify for Big Volume as they contain annual claim records for all physicians submitting to Medicare within the entire United States. Every year, CMS releases the data for a previous year increasing the Big Volume of available data. The datasets contain around 30 attributes each, ranging from provider demographics and the types of procedures to payment amounts and the number of services performed, thus qualifying as Big Variety. Additionally, the Combined dataset used in our study inherently provides Big Variety data, because it combines the three key (but different) Medicare data sources. As CMS is a government program with transparent quality controls and detailed documentation for each dataset, we believe that these datasets are dependable, valid, and representative of all known Medicare provider claims indicating Big Veracity. Through research conducted by our research group and others, it is evident that this data can be used to detect fraudulent behavior giving it Big Value. Furthermore, the LEIE dataset could also be considered as Big Value since it contains the largest known repository of real-world fraudulent medical providers in the United States.

The contributions of this study are twofold. First, we provide detailed discussions on Medicare Big Data processing and exploratory experiments and analyses to show the best learners and datasets for detecting Medicare provider claims fraud. Our unique data processing steps consist of data imputation, determining which variables (dataset features) to keep, transforming the data from the procedure-level to the provider-level through aggregation to match the level of the LEIE dataset for fraud label mapping, and creating the Combined dataset. Note that the fraud labels are

used to assess fraud leveraging historical exclusion information, as well as payments made by Medicare to currently excluded providers. Second, the resulting processed datasets are considered Big Data and thus, for our fraud detection experiments, we employ Spark [26] on top of a Hadoop [27] YARN cluster which can effectively handle these large dataset sizes. For our experiments, the four Medicare datasets were trained and validated using fivefold cross-validation, and the process was repeated ten times. From the Apache Spark 2.3.0 [28] Machine Learning Library, we build the Random Forest (RF), Gradient Tree Boosting (GTB) and Logistic Regression (LR) models, and use the Area under the ROC Curve (AUC) metric to gauge fraud detection performance. We chose these learners, as they and commonly used and provide reasonably good performance, for our exploratory analysis to assess fraud detection performance using Big Data in Medicare. In order to add robustness around the results, we estimate statistical significance with the ANalysis Of VAriance (ANOVA) [29] and Tukey's Honest Significant Difference (HSD) tests [30]. Our results indicate that the Combined dataset with LR resulted in the highest overall AUC with 0.816, while the Part B dataset with LR was the next best with 0.805. Additionally, the Part B dataset had the best results for GBT and RF with both resulting in a 0.796 AUC. The worst fraud detection results were attributed to the DMEPOS dataset, with RF having the lowest overall AUC of 0.708. The results for the Combined dataset using LR, indicate better performance than any individual Medicare dataset; thus, the whole in this case is better than the sum of its parts. This, however, is not the case for RF or GBT with Part B having the highest average AUC. Even so, the Combined dataset showed no statistical difference when compared to the Part B dataset results. Therefore, the high fraud detection results, paired with our assumption that Medicare fraud can be committed in any or all parts of Medicare, demonstrates the potential in using the Combined dataset to successfully detect provider claims fraud across learners. To summarize, the unique contributions of this paper are as follows:

- Detailing Medicare Part B, Part D, and DMEPOS data processing and real-world fraud label mapping.
- Combining the three Medicare big datasets into one Combined dataset to demonstrate high fraud detection performance that takes into account the different key parts of Medicare.
- Exploring fraud detection performance and learner behavior for each of the four big datasets.

The rest of the paper is organized as follows. "Related works" section covers related works, focusing on works employing multiple CMS branches of Medicare. "Datasets" section discusses the different Medicare datasets used, how the data is processed, and the fraud label mapping approach. "Methods" section details the methods used including the learners, performance metric, and hypothesis testing. "Results and discussion" section discusses the results of our experiment. Finally, we conclude and discuss future work in "Conclusion" section.

Related works

There have been a number of studies conducted, by our research group and others, using Public Use Files (PUF) data from CMS in assessing potential fraudulent activities through data mining and other analytics methods. The vast majority of these studies use only Part B data [17, 31–37], neglecting to account for other parts of Medicare when detecting fraudulent behavior. Within the healthcare system, anywhere money is being exchanged, there is an opportunity for a bad actor to manipulate the process and siphon funds, affecting the efficiency and effectiveness of the Medicare healthcare process. There is limited prior information as to where (in the Medicare system) a physician will commit fraud, so choosing a single part of Medicare could miss fraud committed elsewhere. In this study, we focus on the processing and labeling of each Medicare dataset and fraud detection performance. Therefore, we generally limit our discussion in this section to the small body of works attempting to identify fraudulent behavior using multiple CMS datasets. As of this study, we only found two works [38, 39] that fall under that category.

In [38], Branting et al. use the Part B (2012–2014), Part D (2013) and LEIE dataset. They do not specifically mention how they preprocess the data or combine Part B and Part D, but they do take attributes from both Part B and Part D datasets, treating drugs and HCPCS codes in the same way. They matched 12,153 fraudulent physicians using the National Provider Identifier (NPI) [40] with their unique identity-matching algorithm. They decided against distinguishing between LEIE exclusion rules/codes and instead used every listed physician. It is unclear whether the authors accounted for waivers, exclusion start dates or the length of the associated exclusion during their fraud label mapping process. These details are important in reducing redundant and overlapping exclusion labels and for assessing accurate fraud detection performance. Therefore, due to this lack of clarity in the exclusion labeling methodology, the results from their study cannot be reliably reproduced and can be difficult to compare to other research. They developed a method for pinpointing fraudulent behavior by determining the fraud risk through the application of network algorithms from graphs. Due to the highly imbalanced nature of the data, the authors used a 50:50 class distribution, retaining 12,000 excluded providers while randomly selecting 12,000 non-excluded providers. They put forth a few groups of algorithms and determined their fraud detection results based on the real-world fraudulent physicians found in the LEIE dataset. One set of algorithms, which they denote as Behavior–Vector similarity, determines similarity in behavior for real-world fraudulent and non-fraudulent physicians using nominal values such as drug prescriptions and medical procedures. Another group of algorithms makes up their risk propagation, which uses geospatial co-location (such as location of practice) in order to estimate the propagation of risk from fraudulent healthcare providers. An ablation analysis showed that most of this predictive accuracy was the result of features that measure risk propagation through geospatial collocation.

Sadiq et al. [39] use the 2014 CMS Part B, Part D and DMEPOS datasets (using only the provider claims from Florida) in order to find anomalies that possibly point to fraudulent or other interesting behavior. The authors do not go into detail on how they preprocessed the data between these datasets. From their study, we can assume the authors use, at minimum, the following features: NPI, gender, location (state, city, address etc.),

type, service number, average submitted charge amount, the average allowed amount in Medicare and the average standard amount in Medicare. It is also unclear as to whether they used the datasets together or separately or which attributes are used and which are not, making the reproduction of these experiments difficult. The authors determine that when dealing with payment variables, it is best to go state-by-state as each state's data can vary. However, in this paper, we found that good results can be achieved by using Medicare data encompassing the entire U.S. The framework they employ is the Patient Rule Induction Method based bump hunting method, which is an unsupervised approach attempting to determine peak anomalies by spotting spaces of higher modes and masses within the dataset. They explain that by applying their framework, they can characterize the attribute space of the CMS datasets helping to uncover the events provoking financial loss.

We note a number of differences from these two studies [38, 39] including data processing methods, the process for data combining and comparisons made between the three Medicare datasets both individually and combined. We provide a detailed account of the data processing methods for each Medicare dataset as well as the mapping and generation of fraud labels using the LEIE dataset. To the best of our knowledge, this is the first study to compare fraud detection within three different Medicare big datasets, as well as a Combined version of the three primary Medicare datasets, with no other known related studies. Even though our experiments are exploratory in nature, we provide a more complete and comprehensive study, with in-depth data processing details, than what is currently available in this area, using three different learners and four datasets. Additionally, we incorporate all available years in each CMS dataset covering the entire United States, requiring us to incorporate software which can handle such Big Data.

Datasets

In this section, we describe the CMS datasets we use (Part B, Part D and, DMEPOS). Furthermore, the data processing methodology used to create each dataset, including processing, fraud label mapping between the Medicare datasets and the LEIE, and one-hot encoding for categorical variables is discussed. The information within each dataset is based on CMS's administrative claims data for Medicare beneficiaries enrolled in the Fee-For-Service program. Note, this data does not take into account any claims submitted through the Medicare Advantage program [41]. Since CMS records all claims information after payments are made [42–44], we assume the Medicare data is already cleansed and is correct. Note that NPI is not used in the data mining step, but rather for aggregation and identification. Additionally, for each dataset, we added a year variable which is also used for aggregation and identification.

Medicare dataset descriptions

Part B

The Part B dataset provides claims information for each procedure a physician performs within a given year. Currently, this dataset is available on the CMS website for the 2012 through 2015 calendar years (with 2015 being released in 2017) [45]. Physicians are identified using their unique NPI [40], while procedures are labeled by their Healthcare

Common Procedure Coding System (HCPCS) code [46]. Other claims information includes average payments and charges, the number of procedures performed and medical specialty (also known as provider type). CMS decided to aggregate Part B data over: (1) NPI of the performing provider, (2) HCPCS code for the procedure or service performed, and (3) the place of service which is either a facility (F) or non-facility (O), such as a hospital or office, respectively. Each row, in the dataset, includes a physician's NPI, provider type, one HCPCS code split by place of service along with specific information corresponding to this breakdown (i.e. claim counts) and other non-changing attributes (i.e. gender). We have found that in practice, physicians perform the same procedure (HCPCS code) at both a facility and their office, as well as a few physicians that practice under multiple provider types (specialties) such as Internal Medicine and Cardiology. Therefore, for each physician, there are as many rows as unique combinations of NPI, Provider Type, HCPCS code and place of service and thus Part B data can be considered to provide procedure-level information. Table 1 provides an example of one physician with NPI = 1649387770 sampled from the 2015 Part B dataset.

Part D

The Part D dataset provides information pertaining to the prescription drugs they administer under the Medicare Part D Prescription Drug Program within a given year. Currently, this data is available on the CMS website for the 2013 through 2015 calendar years (with 2015 being released in 2017) [47]. Physicians are identified using their unique NPI within the data while each drug is labeled by their brand and generic name. Other information includes average payments and charges, variables describing the drug quantity prescribed and medical specialty. CMS decided to aggregate the Part D data over: (1) the NPI of the prescriber, and (2) the drug name (brand name in the case of trademarked drugs) and generic name. Each row in the Part D dataset lists a physician's NPI, provider type and drug name along with specific information corresponding to this breakdown (i.e. claim counts) and other static attributes (i.e. gender). Same as with Part B, we found a few physicians that practice under multiple specialties, such as Internal Medicine and Cardiology. Therefore, for each physician, there are as many rows as unique combinations of NPI, Provider Type, drug name and generic name and thus, Part D data can be considered to provide procedure-level information. In order to protect the privacy of Medicare beneficiaries, any aggregated records, derived from 10 or fewer claims, are excluded from the Part D data. Table 2 provides an example of one physician with NPI = 1649387770 sampled from the 2015 Part D dataset.

Table 1 Sample of the Part B dataset

Npi	...	Provider_type	...	Place_of _service	Hcpcs _code	...	Line_srvc _cnt	...	Average_ submitted _chrg_amt	...
1649387770	...	Ophthalmology	...	O	66821	...	28	...	1200	...
1649387770	...	Ophthalmology	...	F	66984	...	154	...	2400	...
1649387770	...	Ophthalmology	...	O	67820	...	45	...	105	...
1649387770	...	Ophthalmology	...	O	76514	...	11	...	80	...
1649387770	...	Ophthalmology	...	O	92004	...	205	...	175	...

Table 2 Sample of Part D dataset

Npi	...	Provider_type	...	Drug_name	Total_drug _cost	Total_claim _count_ ge65	Ge65 _suppress _flag	...
1649387770	...	Ophthalmology	...	ALPHAGAN P	11811.27	57	NA	...
1649387770	...	Ophthalmology	...	AZASITE	3410.56	25	NA	...
1649387770	...	Ophthalmology	...	AZOPT	8336.27	27	NA	...
1649387770	...	Ophthalmology	...	BRIMONIDINE TARTRATE	1769.25	12	NA	...
1649387770	...	Ophthalmology	...	COMBIGAN	25434.18	127	NA	...

Table 3 Sample of DMEPOS

Referring_npi	...	Referring _provider_ type	...	Hcpcs_code	...	Supplier _rental _indicator	Number_of _supplier _claims	Avg_supplier _submitted _charge	...
1649387770	...	Ophthalmology	...	V2020	...	N	44	67.4	...
1649387770	...	Ophthalmology	...	V2203	...	N	21	66.0	...
1649387770	...	Ophthalmology	...	V2303	...	N	18	87.5	...

DMEPOS

The DMEPOS dataset provides claims information about Medical Equipment, Prosthetics, Orthotics and Supplies that physicians referred patients to either purchase or rent from a supplier within a given year. Note, this dataset is based on supplier's claims submitted to Medicare while the physician's role is referring the patient to the supplier. Currently this data is available on the CMS website for 2013 through 2015 calendar years (with 2015 being released in 2017) [48]. Physicians are identified using their unique NPI within the data while products are labeled by their HCPCS code. Other claims information includes average payments and charges, the number of services/products rented or sold and medical specialty (also known as provider type). CMS decided to aggregate Part B data over: (1) NPI of the performing provider, (2) HCPCS code for the procedure or service performed by the DMEPOS supplier, and (3) the supplier rental indicator (value of either 'Y' or 'N') derived from DMEPOS supplier claims (according to CMS documentation). Each row provides a physician's NPI, provider type, one HCPCS code split by rental or non-rental with specific information corresponding to this breakdown (i.e. number of supplier claims) and other non-changing attributes (i.e. gender). We have found that some physicians place referrals for the same DMEPOS equipment, or HCPCS code, as both rental and non-rental as well as a few physicians that practice under multiple specialties such as Internal Medicine and Cardiology. Therefore, for each physician, there are as many rows as unique combinations of NPI, Provider Type, HCPCS code and rental status, and thus the DMEPOS data also can be considered to provide procedure-level information. Table 3 provides an example of one physician with NPI = 1649387770 from the 2015 DMEPOS dataset.

Table 4 Sample of LEIE

Specialty	...	Npi	...	Excltype	Excldate	...
GENERAL PRACTICE/FP	...	0	...	1128b6	19770701	...
EMPLOYEE	...	0	...	1128b6	19780124	...
GENERAL PRACTICE	...	1003016742	...	1128a1	20170720	...
NURSE/NURSES AIDE	...	1003011644	...	1128b4	20091220	...

Table 5 LEIE rules involving fraud

Rule number	Description
1128(a)(1)	Conviction of program-related crimes
1128(a)(2)	Conviction relating to patient abuse or neglect
1128(a)(3)	Felony conviction relating to health care fraud
1128(b)(4)	License revocation or suspension
1128(b)(7)	Fraud, kickbacks and other prohibited activities
1128(c)(3)(g)(i)	Conviction of two mandatory exclusion offenses 10 years
1128(c)(3)(g)(ii)	Conviction of 3 mandatory exclusion offenses indefinite

LEIE

In order to accurately assess fraud detection performance as it appears in real-world practice, we require a data source that contains physicians that have committed real-world fraud. Therefore, we employ the List of Excluded Individuals and Entities (LEIE) [19], which contains the following information: reason for exclusion, date of exclusion and reinstate/waiver date for all current physicians found unsuited to practice medicine and thus excluded from practicing in the United States for a given period of time. This dataset was established and is maintained monthly by the Office of Inspector General (OIG) [49] in accordance with Sections 1128 and 1156 of the Social Security Act [50]. The OIG has authority to exclude individuals and entities from federally funded healthcare programs, such as Medicare. Unfortunately, the LEIE is not all-inclusive where 38% of providers with fraud convictions continue to practice medicine and 21% were not suspended from medical practice despite their convictions [51]. Moreover, the LEIE dataset only contains the NPI values for a small percentage of physicians and entities. An example of four different physicians and how they are portrayed within the LEIE is shown in Table 4, where any physician without a listed NPI has a value of 0.

The LEIE is aggregated at the provider-level and does not have specific information regarding procedures, drugs or equipment related to fraudulent activities. There are different categories of exclusions, based on severity of offense, described by various rule numbers. We do not use all exclusions, but rather filter the excluded providers by selected rules indicating fraud was committed [34]. Table 5 gives the codes that correspond to fraudulent provider exclusions and the length of mandatory exclusion. We have determined that any behavior prior to and during a physician's "end of exclusion date" constitutes fraud.

Data processing

For each dataset (Part B, Part D and DMEPOS), we combined the information for all available calendar years [52]. For our research, Part B was available for 2012 through 2015, while Part D and DMEPOS were available for 2013 through 2015. For Part B and DMEPOS, the first step was removing all attributes not present in each available year. The Part D dataset had the same attributes in all available years. For Part B, we removed the standard deviation variables from 2012 and 2013 and standardized payment variables from 2014 and 2015 as they were not available in the other years. For DMEPOS, we removed a standard deviation variable from 2014 and 2015 as it was not available in 2013. For all three datasets, we removed all instances that either were missing both NPI and HCPCS/drug name values or had an invalid NPI (i.e. NPI = 0000000000). For Part B, we filtered out all instances with HCPCS codes referring to prescriptions. These prescription-related codes are not actual medical procedures, but instead are for specific services listed on the Medicare Part B Drug Average Sales Price file [11]. Keeping these instances would muddy the results as the line_srvc_cnt feature in these cases represents weight or volume of a drug, rather than simply quantifying procedure counts.

For this study, we are only interested in particular attributes from each dataset in order to provide a solid basis for our experiments and analyses. For the Part B dataset, we kept eight features while removing the other twenty-two. For the Part D dataset, we kept seven and removed the other fourteen. For the DMEPOS dataset we kept nine and removed the other nineteen. The excluded attributes provide no specific information on the claims, drugs administered, or referrals, but rather encompass provider-related information, such as location and name, as well as redundant variables like text descriptions which can be represented by using the variables containing the procedure or drug codes. For Part D, we also did not include variables that provided count and payment information for patients 65 or older as this information is encompassed in the kept variables. In this case, the claim count variable (total_claim_count) contains counts for all ages to include patients 65 or older. Tables 6, 7 and 8 detail the features we chose from the datasets, including a description and feature type (numerical or categorical) along with the exclusion attribute (fraud label) derived from the LEIE.

The data processing steps are similar for Part B, Part D and DMEPOS. All three unaltered datasets are originally at the HCPCS or procedure level, meaning they were aggregated by NPI and HCPCS/drug. To meet our needs of mapping fraud labels using the

Table 6 Description of features chosen from the Part B dataset

Feature	Description	Type
Npi	Unique provider identification number	Categorical
Provider_type	Medical provider's specialty (or practice)	Categorical
Nppes_provider_gender	Provider's gender	Categorical
Line_srvc_cnt	Number of procedures/services the provider performed	Numerical
Bene_unique_cnt	Number of distinct Medicare beneficiaries receiving the service	Numerical
Bene_day_srvc_cnt	Number of distinct Medicare beneficiary/per day services	Numerical
Average_submitted_chrg_amt	Average of the charges that the provider submitted for the service	Numerical
Average_medicare_payment_amt	Average payment made to a provider per claim for the service	Numerical
Exclusion	Fraud labels from the LEIE dataset	Categorical

Table 7 Description of features chosen from the Part D dataset

Feature	Description	Type
Npi	Unique provider identification number	Categorical
Specialty_description	Medical provider's specialty (or practice)	Categorical
Bene_count	Number of distinct Medicare beneficiaries receiving the drug	Numerical
Total_claim_count	Number of drug the provider administered	Numerical
Total_30_day_fill_count	Number of standardized 30-day fills	Numerical
Total_day_supply	Number of day's supply	Numerical
Total_drug_cost	Cost paid for all associated claims	Numerical
Exclusion	Fraud labels from the LEIE dataset	Categorical

Table 8 Description of features chosen from the DMEPOS dataset

Feature	Description	Type
Referring_npi	Unique provider identification number	Categorical
Referring_provider_type	Medical provider's specialty (or practice)	Categorical
Referring_provider_gender	Provider's gender	Categorical
Number_of_suppliers	Number of suppliers used by provider	Numerical
Number_of_supplier_beneficiaries	Number of beneficiaries associated by the supplier	Numerical
Number_of_supplier_claims	Number of claims submitted by a supplier from a referring order	Numerical
Number_of_supplier_services	Number of services/products rendered by a supplier	Numerical
Avg_supplier_submitted_charge	Average payment submitted by a supplier	Numerical
Avg_supplier_medicare_pmt_amt	Average payment awarded to suppliers	Numerical
Exclusion	Fraud labels from the LEIE dataset	Categorical

LEIE, we reorient each dataset, aggregating to the provider-level where all information is grouped by and aggregated over each NPI (and other specific features). For Part B, the aggregating process consists of grouping the data by NPI, provider type, gender and year, aggregating over HCPCS and place of service. Part D was grouped by NPI, provider type and year aggregating over drugs. DMEPOS was grouped by NPI, provider type, gender and year, aggregating over HCPCS and rental status. For the Part D and DMEPOS datasets, their beneficiary counts are suppressed to 0 if originally below 11, and in response we imputed the value of 5 as recommended by CMS.

In an effort to bypass information loss due to aggregating these datasets, we generated six numeric features for each chosen numeric feature outlined in the previous subsection for each dataset ("Medicare dataset descriptions" section). Therefore, for each numeric value, per year, in each dataset, we replace the original numeric variables with the aggregated mean, sum, median, standard deviation, minimum and maximum values, creating six new features for each original numeric feature. The resulting features are all complete except for standard deviation which contains NA values. These NA values are generated when a physician has performed/prescribed a HCPCS/drug once in a given year. Therefore, the population standard deviation for one unique instance is 0, and thus we replace all NA values with 0 representing that this single instance has no variability in that particular year. Two other features included are the categorical features: provider type and gender (Part D does not contain a gender variable).

Combined dataset

The Combined dataset is created after processing Part B, Part D, and the DMEPOS datasets, containing all the attributes from each, along with the fraud labels derived from the LEIE. The combining process involves a join operation on NPI, provider type, and year. Due to there not being a gender variable present in the Part D data, we did not include this variable in the join operation conditions and used the gender labels from Part B while removing the gender labels gathered from the DMEPOS dataset after joining. In combining these datasets, we are limited to those physicians who have participated in all three parts of Medicare. Even so, this Combined dataset has a larger and more encompassing base of attributes for applying data mining algorithms to detect fraudulent behavior, as demonstrated in our study.

Fraud labeling

For all four datasets, we use the LEIE dataset for generating fraud labels, where only physicians within are considered fraudulent, otherwise they are considered non-fraudulent. In order to obtain exact matches between the Medicare datasets and the LEIE, we determined that the NPI value is the only way to match physicians exactly, assuring our data the utmost reliability. The LEIE gives specific dates (month/day/year) for when the exclusion starts and the length of the exclusion period, where we use only month/year (no rounding within a month, i.e. May 1st through 31st is considered May). For example, if a provider breaks rule number 1128(a)(3) ('felony conviction due to healthcare fraud') carrying a minimum exclusion period of 5 years beginning February 2010, then the end of the exclusion period would be February 2015. Note that we used the earliest date between the exclusion end date (based on minimum exclusion period summed with start date), waiver, and reinstatement date. Therefore, continuing this example, if there is also a waiver date listed as October 2014 and a reinstatement date of December 2014, the exclusion period would be between February 2010 and October 2014. This accounts for providers that may still be in their exclusion period but received a waiver or reinstatement to use Medicare, thus no longer considered fraudulent on or after this waiver or reinstatement date.

Contrary to the LEIE data, the Medicare datasets are released annually where all data is provided for each given year. In order to best handle the disparity between the annual and monthly dates, we round the new exclusion end date to the nearest year based on the month. If the end exclusion month is greater than 6 (majority of the year), then the exclusion end year is increased to the following year; otherwise, the current year is used. We do not want a physician to be considered fraudulent during a year unless more than half that year is before their exclusion end date. Continuing the above example, we determined that the end exclusion date was October 2014, therefore since October is the tenth month and 10 is greater than 6, the end exclusion year would be rounded up to 2015. Therefore, translating this to the Medicare data, any activity in 2014 or earlier would be considered fraudulent when creating fraud labels. For further clarification, if the waiver date would have been March 2014, the end exclusion year would be 2014 and only activity from 2013 or earlier would be labeled fraudulent.

Table 9 Distribution of fraud labels

Dataset	Non-fraudulent	Fraudulent	% Fraudulent
Part B	3,691,146	1409	0.038
Part D	2,098,715	1018	0.048
DMEPOS	862,792	635	0.074
Combined	759,267	473	0.062

The LEIE dataset is joined to all four datasets based on NPI. We create an exclusion feature which is the final categorical attribute discussed in previous sections, which indicates either fraud or non-fraud instances. Any physician practicing within a year prior to their exclusion end year is labeled fraudulent. With an exclusion year of 2015, from the physician in our previous example, for Part B, the years 2012 through 2014 would be labeled fraudulent, while for Part D, DMEPOS, and the Combined datasets, 2013 and 2014 would be marked fraudulent (as 2012 is not available for these datasets). Through this process, we are accounting for two types of fraudulent behavior: (1) actual fraudulent behavior, and (2) payments made by Medicare based on submissions from excluded providers, where both drain funds from Medicare inappropriately. For the former, we assume any activity before being caught/excluded is fraudulent behavior. We also include the latter as fraud because, according to the False Claims Act (FCA), this is a form of fraudulent behavior [53]. The final four datasets include all known excluded providers marked via the categorical exclusion feature. Table 9 shows the distribution of fraud to non-fraud within all four datasets. All four datasets are considered highly imbalanced, ranging between 0.038% and 0.074% of instances being labeled as fraud. In this exploratory work, we do not apply techniques to address class imbalance [54–56], leaving this as future work.

One-hot encoding

In order to build our models with a combination of numerical and categorical features, we employ one-hot encoding, transforming the categorical features. For example, one-hot encoding gender would first consist of generating extra features equaling the number of options, in this case two (male and female). If the physician is male, the new male feature would be assigned a 1 and the female feature would be 0; while for female, the male would be assigned a 0 and the female assigned a 1. If the original gender feature is missing then both male and female are assigned a 0. This process is done for all four datasets for gender and provider type/specialty. Table 10 summarizes all four datasets after data processing and after the categorical features have been one-hot encoded. Note that NPI is not used for building models and is removed from each dataset after this step.

Table 10 Summary of Medicare datasets

	Part B		Part D		DMEPOS		Combined	
	Instances	Features	Instances	Features	Instances	Features	Instances	Features
After processing and fraud labeling	3,692,555	35	2,099,733	34	863,427	41	759,740	102
After one-hot encoding	3,692,555	126	2,099,733	126	863,427	145	759,740	173

Methods

Learners

For running and validating models, we used Spark on top of a Hadoop Yarn cluster due to the Big Volume of the datasets. We used three classification models available in the Apache Spark 2.3.0 Machine Learning Library: Logistic Regression, Gradient Boosted Trees and Random Forest. In this section, we briefly describe each learner and note any configuration changes that differ from the default settings.

Logistic Regression (LR) [57] predicts probabilities for which class a categorical dependent variable belongs to by using a set of independent variables employing a logistic function. LR uses a sigmoidal (logistic) function to generate values that can be interpreted as class probabilities. LR is similar to linear regression but uses a different hypothesis class to predict class membership [58–61]. The bound matrix was set to match the shape of the data (number of classes and features) so the algorithm knows the number of classes and features the dataset contains. The bound vector size is equal to 1 for binomial regression, and no thresholds are set for binary classification.

Random Forest (RF) [62, 63] is an ensemble learning method that generates a large number of trees. The class value appearing most frequently among these trees is the class predicted as output from the model. As an ensemble learning method, RF is an aggregation of various tree predictors. Each tree within the forest is dependent upon the values dictated by a random vector that is independently sampled and where each tree is equally distributed among the forest [60, 64]. The RF ensemble inserts randomness into the training process which can minimize overfitting and is fairly robust to imbalanced data [65, 66]. We build each RF learner with 100 trees as our research group has found little to no benefit using more trees. The parameter that caches node IDs for each instance, was set to true and the maximum memory parameter was set to 1024 MB in order to minimize training time. The setting that manipulates the number of features to consider for splits at each tree node was set to one-third, since this setting provided better results upon initial investigation. The maximum bins parameter, which is the max number of bins for discretizing continuous features, is set to 2 because we no longer have categorical features since they were converted using one-hot encoding.

Gradient Boosted Trees (GBT) [62, 63] is another ensemble of decision trees. Unlike RF, GBT trains each decision tree one at a time in order to minimize loss determined by the algorithm's loss function. During each iteration, the current ensemble is used to

predict the class for each instance in the training data. The predicted values are evaluated with the actual values allowing the algorithm to pinpoint and correct previously mislabeled instances. The parameter that caches node IDs for each instance, was set to true and the maximum memory parameter was set to 1024 MB to minimize training time.

Performance metric

In assessing Medicare fraud, we are presented with a two-class classification problem where a physician is either fraudulent or non-fraudulent. In our study, the positive class, or class of interest, is fraud and the negative class is non-fraud. Spark presented us with a confusion matrix for each model and is commonly used to assess the performance of learners. Confusion matrices provide counts comparing actual counts against predicted counts. From the resultant matrices, we employ AUC [67, 68] to measure fraud detection performance. AUC is the Area under the Receiver Operating Characteristic (ROC) curve, where ROC is the comparison between false positive (fall-out) and true positive (recall). Recall is calculated by $\frac{TP}{TP+FN}$ and fall-out is calculated by $\frac{FP}{FP+TN}$. The definitions for TP, TN, FP and FN, which can be directly calculated from the confusion matrix are as follows:

- True positive (TP): number of actual positive instances correctly predicted as positive.
- True negative (TN): number of actual negative instances correctly predicted as negative.
- False positive (FP): number of negative instances incorrectly classified as positive.
- False negative (FN): number of positive instances incorrectly assigned as negative.

The AUC curve is an encompassing evaluation of a learner as it depicts performance across all decision thresholds. The AUC results in a single value ranging from 0 to 1, where a perfect classifier results in an AUC of 1, an AUC of 0.5 is equivalent to random guessing and less than 0.5 demonstrates bias towards a given class. AUC has been found to be effective for class imbalance [69].

Cross-validation

We employ stratified k-fold cross-validation in evaluating our models, where $k = 5$. Stratification ensures all folds have class representation matching the ratio of the original data, which is important when dealing with largely imbalanced data. The training data is evenly divided into fivefold where fourfold will be used for training the model and the remaining fold tests the model. This process is repeated 5 times allowing each fold an opportunity as the test fold, ensuring the entire dataset is fully leveraged being used in training and validation. Spark will automatically create different folds each time the learner is run, and to validate our results we ran each model 10 times for each learner/dataset pair. The use of repeats helps to reduce bias due to bad random draws when creating the folds where the final performance for every presented result is the average over all 10 repeats.

Significance testing

In order to provide additional rigor around our AUC performance results, we use hypothesis testing to show the statistical significance of the Medicare fraud detection results. Both ANOVA and post hoc analysis via Tukey's HSD tests are used in our study. ANOVA is a statistical test determining whether the means of several groups (or factors) are equal. Tukey's HSD test determines factor means that are significantly different from each other. This test compares all possible pairs of means using a method similar to a t-test, where statistically significant differences are grouped by assigning different letter combinations (e.g. group a is significantly different than group b).

Results and discussion

This section discusses the results of our study, assessing dataset and learner performance for Medicare fraud detection. The practices of individual physicians are unique, where a given physician might only submit claims to Medicare through Part B, Part D, DMEPOS, or to all three. Therefore, we show learner performance in relation to each of the Medicare datasets to establish the best fraud detection combinations. In Table 11, we show the AUC results for each dataset and learner combination. The italicized values depict the highest AUC scores per dataset, whereas the underlined values are the highest per learner. LR produces the two highest overall AUC scores for the Combined dataset with 0.816 and Part B with 0.805. The Combined dataset has the best overall AUC, but the Part B dataset shows the lowest variation in fraud detection performance across learners, which includes having the highest AUC scores for GBT and RF. The Part D and DMEPOS datasets have the lowest AUC values for all three learners, but show improvement when using LR and GBT compared to RF.

The favorable results using LR with each of the datasets may be due to the squared-error loss function with the application of L2 regularization, also known as Ridge Regression, penalizing large coefficients and improving the generalization performance, making LR fairly robust to noise and overfitting. Even though LR performs well on the Part B and Combined datasets, additional testing is required to determine whether the Part D and DMEPOS datasets have particular characteristics contributing to their lower fraud detection performance. The poor performance of the tree-based methods, particularly RF, may be due to the lack of independence between individual trees or the high cardinality of the categorical variables. The Combined dataset contains features across the three parts of Medicare creating a robust pool of attributes, presumably allowing for better model generalization and overall fraud detection performance. In particular, the Combined dataset using LR has the highest AUC with better performance versus each of its individual Medicare parts. This is not

Table 11 Learner AUC results by dataset

Dataset	Logistic Regression	Gradient Boosted Trees	Random Forest
Combined	*0.81554*	0.79047	0.79383
Part B	*0.80516*	0.79569	0.79604
Part D	*0.78164*	0.74851	0.70888
DMEPOS	*0.74063*	0.73129	0.70756

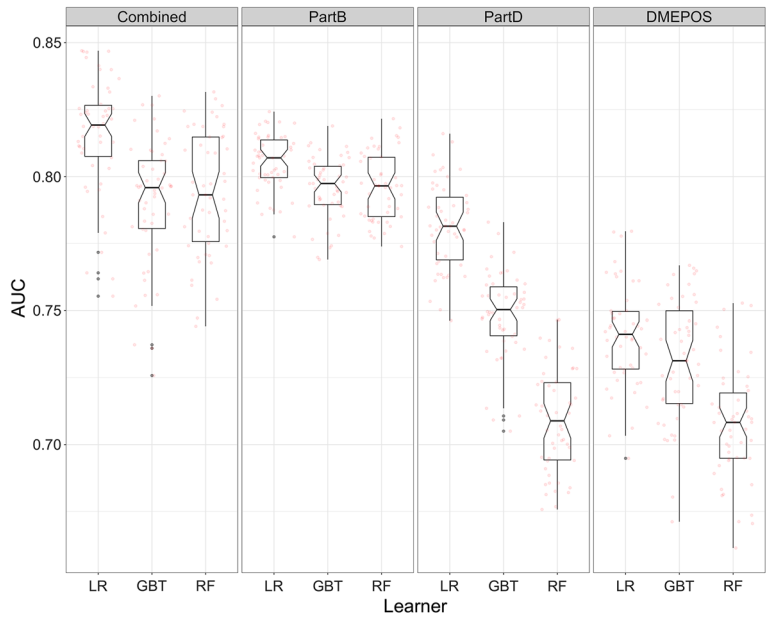

Fig. 1 AUC values for 10 runs of fivefold cross-validation for each learner and dataset combination

Table 12 Two-factor ANOVA test results

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dataset	3	0.6257	0.20855	594.15	< 2e−16
Learner	2	0.1174	0.05868	167.17	< 2e−16
Dataset:Learner	6	0.0658	0.01097	31.26	< 2e−16
Residuals	588	0.2064	0.00035	–	–

the case with RF or GBT, with Part B indicating the highest AUC scores. Interestingly, the Part B dataset has the lowest variability across the learners and within each individual learner, which could be due, in part, to having the largest number of fraud labels. The Part D and DMEPOS datasets not only show poor learner performance, but exhibit generally higher AUC variability across individual learners. This could indicate possible adverse effects of high class imbalance or less discriminatory power in the selected features. With regards to our above discussions, Fig. 1 shows a box plot of our experimental results over all 50 AUC values from the ten runs of fivefold cross-validation for each dataset/learner pair.

Table 12 presents the results for the two-factor ANOVA test over each Dataset and Learner, as well as their interaction (Dataset:Learner). The ANOVA test shows that these factors and their interactions are statistically significant at a 95% confidence interval. In order to determine statistical groupings, we perform a Tukey's HSD test on the results for the Medicare datasets, which corroborates the high performance of the LR learner and the Combined dataset for Medicare fraud detection (as seen in Table 11).

In Table 13, the results for each learner across all datasets show that LR is significantly better than GBT and RF. Moreover, LR and GBT have similar AUC variability, but LR has the highest minimum and maximum AUC scores which, again, substantiate the

Table 13 Two-factor Tukey's HSD learner results over all datasets

Learner	Group	AUC	sd	r	Min	Max
Logistic Regression	a	0.78574	0.03369	200	0.69487	0.847
Gradient Boosted Trees	b	0.76649	0.03343	200	0.67119	0.83013
Random Forest	c	0.75158	0.04753	200	0.66138	0.83161

Table 14 Two-factor Tukey's HSD dataset results over all learners

Dataset	Group	AUC	sd	r	Min	Max
Combined	a	0.79995	0.02549	150	0.7258	0.847
Part B	a	0.79896	0.0123	150	0.769	0.82425
Part D	b	0.74634	0.03443	150	0.67576	0.81602
DMEPOS	c	0.72649	0.02506	150	0.66138	0.77957

good performance of LR for each dataset. Table 14 summarizes the significance of dataset performance across each learner. We notice that the Combined and Part B datasets show significantly better performance than either the Part D or DMEPOS datasets, and that the DMEPOS dataset is significantly worse than Part D dataset. Since the Part B and Combined results are not significantly different, we consider the Combined dataset preferable for general fraud detection since we do not necessarily know beforehand exactly which part of the Medicare system a physician/provider will target any fraudulent behavior (e.g. medical procedures/services, drug submissions, or prosthetic rental). With the Combined dataset, we have a larger web for monitoring fraudulent behavior as opposed to monitoring only one part of Medicare for a given healthcare provider. Additionally, the Combined dataset with LR provides the only results where the Combined dataset produces the best performance, greater than the results for the individual Medicare datasets. Therefore, based on these exploratory performance results, we demonstrate that when a physician has participated in Part B, Part D, and DMEPOS, the Combined dataset, using LR, indicates the best overall fraud detection performance.

Conclusion

The importance of reducing Medicare fraud, in particular for individuals 65 and older, is paramount in the United States as the elderly population continues to grow. Medicare is necessary for many citizens, and therefore, the importance placed on quality research into fraud detection to keep healthcare costs fair and reasonable. CMS has made available several Big Data Medicare claims datasets for public use over an ever-increasing number of years. Throughout this work, we provide a unique approach (combining multiple Medicare datasets and leverage state-of-the-art Big Data processing and machine learning approaches) for determining the fraud detection capabilities of three Medicare datasets, individually and combined, using three learners, against real-world fraudulent physicians and other medical providers taken from the LEIE dataset.

We present our methods for processing each dataset from CMS, the Combined dataset, as well as the mapping of provider fraud labels. We ran experiments on all four datasets: Part B, Part D, DMEPOS, and Combined. Each dataset was considered Big Data,

requiring us to employ Spark on top of a Hadoop YARN cluster for running and validating our models. Each dataset was trained and evaluated using three learners: Random Forest, Gradient Boosted Trees and Logistic Regression. The Combined dataset had the best overall fraud detection performance with an AUC of 0.816 using LR, indicating better performance than each of its individual Medicare parts, and scored similarly to Part B with no significant difference in average AUC. The DMEPOS dataset had the lowest overall results for all learners. Therefore, from these experimental findings and observations, coupled with the notion that a physician/provider can commit fraud using any part of Medicare, we show that using the Combined dataset with LR provides the best overall fraud detection performance. Future work will include employing data sampling techniques to combat the imbalanced nature of known fraud events in evaluating the different Medicare datasets.

Authors' contributions
MH and RAB performed the primary literature review, experimentation and analysis for this work, and also drafted the manuscript. TMK worked with MH to develop the article's framework and focus. TMK introduced this topic to MH, and to complete and finalize this work. All authors read and approved the final manuscript.

Acknowledgements
We would like to thank the reviewers in the Data Mining and Machine Learning Laboratory at Florida Atlantic University. Additionally, we acknowledge partial support by the NSF (CNS-1427536). Opinions, findings, conclusions, or recommendations in this paper are solely of the authors' and do not reflect the views of the NSF.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable.

Publisher's Note

Received: 27 June 2018 Accepted: 21 August 2018
Published online: 04 September 2018

References

1. U.S. Government, U.S. Centers for Medicare & Medicaid Services. The Official U.S. Government Site for Medicare. https://www.medicare.gov/. Accessed 21 Jan 2017.
2. Feldstein M. Balancing the goals of health care provision and financing. Health Affairs. 2006;25(6):1603–11.
3. Administration for Community Living: Profile of older Americans. 2015. http://www.aoa.acl.gov/Aging_Statistics/Profile/2015/. Accessed 2015.
4. Dieleman JL, Squires E, Bui AL, Campbell M, Chapin A, Hamavid H, Horst C, Li Z, Matyasz T, Reynolds A. Factors associated with increases in us health care spending, 1996–2013. Jama. 2017;318(17):1668–78.
5. Medicare Fraud Strike Force. Office of Inspector General. https://www.oig.hhs.gov/fraud/strike-force/. Accessed 9 Feb 2017.
6. Aetna. The facts about rising health care costs. http://www.aetna.com/health-reform-connection/aetnasvision/facts-about-costs.html. Accessed 2015.
7. Morris L. Combating fraud in health care: an essential component of any cost containment strategy. Health Affairs. 2009;28(5):1351–6.
8. CMS: Center for Medicare and Medicaid Services. https://www.cms.gov/. Accessed 2017.
9. Medicare: US Medicare Program. https://www.medicare.gov. Accessed 2017.
10. Henry J. Kaiser family foundation: Medicare advantage. https://www.kff.org/medicare/fact-sheet/medicareadvantage/. Accessed 2017.

11. CMS: Research, Statistics, Data, and Systems. https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems.html. Accessed 18 Nov 2015.
12. Medicare.gov. What's medicare. https://www.medicare.gov/sign-up-change-plans/decide-how-to-get-medicare/whats-medicare/what-is-medicare.html. Accessed 2017.
13. Bauder RA, Khoshgoftaar TM, Seliya N. A survey on the state of healthcare upcoding fraud analysis and detection. Health Serv Outcomes Res Methodol. 2017;17(1):31–55.
14. Rashidian A, Joudaki H, Vian T. No evidence of the effect of the interventions to combat health care fraud and abuse: a systematic review of literature. PLoS ONE. 2012;7(8):41988.
15. Li J, Huang K-Y, Jin J, Shi J. A survey on statistical methods for health care fraud detection. Health Care Manag Sci. 2008;11(3):275–87.
16. Santa Clara. In: Conjunction with the IEEE international conference on BigData: Big Data in bioinformatics and health care informatics. 2013. http://www.ittc.ku.edu/texttildelow jhuan/BBH/ Accessed 12 Apr 2017.
17. Feldman K, Chawla NV. Does medical school training relate to practice? evidence from big data. Big Data. 2015;3(2):103–13.
18. Centers for Medicare & Medicaid Services. Medicare fraud & abuse: prevention, detection, and reporting. 2015. https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/fraud_and_abuse.pdf. Accessed 2017.
19. LEIE: Office of Inspector General LEIE Downloadable Databases. https://oig.hhs.gov/exclusions/authorities.asp. Accessed 14 Jul 2016.
20. Ohlhorst FJ. Big Data analytics: turning Big Data into big money. New York: Wiley; 2012.
21. John Walker S. Big data: a revolution that will transform how we live, work, and think. New York: Taylor & Francis; 2014.
22. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: the next frontier for innovation, competition, and productivity. New York: McKinsey Global Institute; 2011.
23. McAfee A, Brynjolfsson E, Davenport TH, Patil D, Barton D. Big Data: the management revolution. Harvard Bus Rev. 2012;90(10):60–8.
24. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data. 2014;1(1):2.
25. Demchenko Y, Zhao Z, Grosso P, Wibisono A, De Laat C. Addressing big data challenges for scientific data infrastructure. In: 2012 IEEE 4th international conference on cloud computing technology and science (CloudCom). 2012. p. 614–7.
26. Apache: Apache Spark. http://spark.apache.org/. Accessed 24 May 2018.
27. Apache: Apache Hadoop. http://hadoop.apache.org/. Accessed 24 May 2018.
28. Apache: Apache Spark. http://spark.apache.org/news/spark-2-3-0-released.html. Accessed 24 May 2018.
29. Gelman A. Analysis of variance—why it is more important than ever. Ann Stat. 2005;33(1):1–53.
30. Tukey JW. Comparing individual means in the analysis of variance. Biometrics. 1949;5:99–114.
31. Herland M, Bauder RA, Khoshgoftaar TM. Medical provider specialty predictions for the detection of anomalous Medicare insurance claims. In: 2017 IEEE 18th international conference information reuse and integration (IRI). 2017. p. 579–88
32. Bauder RA, Khoshgoftaar TM, Richter A, Herland M. Predicting medical provider specialties to detect anomalous insurance claims. In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI). 2016. p. 784–90.
33. Bauder RA, Khoshgoftaar TM. A probabilistic programming approach for outlier detection in healthcare claims. In: 2016 15th IEEE international conference on machine learning and applications (ICMLA). 2016. p. 347–54.
34. Bauder RA, Khoshgoftaar TM. A novel method for fraudulent medicare claims detection from expected payment deviations (application paper). In: 2016 IEEE 17th international conference on information reuse and integration (IRI). 2016. p. 11–9.
35. Bauder RA, Khoshgoftaar TM. The detection of medicare fraud using machine learning methods with excluded provider labels. In: FLAIRS conference. 2018. p. 404–9.
36. Chandola V, Sukumar SR, Schryver JC. Knowledge discovery from massive healthcare claims data. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2013. p. 1312–20.
37. Khurjekar N, Chou C-A, Khasawneh MT. Detection of fraudulent claims using hierarchical cluster analysis. In: Proceedings IIE annual conference. Institute of Industrial and Systems Engineers (IISE). 2015. p. 2388.
38. Branting LK, Reeder F, Gold J, Champney T. Graph analytics for healthcare fraud risk estimation. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). 2016. p. 845–51.
39. Sadiq S, Tao Y, Yan Y, Shyu M-L. Mining anomalies in medicare big data using patient rule induction method. In: 2017 IEEE third international conference on multimedia Big Data (BigMM). 2017. p. 185–92.
40. CMS: National Provider Identifier Standard (NPI). https://www.cms.gov/Regulations-and-Guidance/Administrative-Simplification/NationalProvIdentStand/. Accessed 8 Nov 2016.
41. US, Medicare Payment Advisory Commission. Report to the Congress, Medicare Payment Policy. Washington D.C.: Medicare Payment Advisory Commission; 2007.
42. CMS Office of Enterprise Data and Analytics. Medicare Fee-For-Service Provider Utilization & Payment Data Physician and Other Supplier. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf. Accessed 9 Apr 2018.
43. CMS Office of Enterprise Data and Analytics. Medicare Fee-For Service Provider Utilization & Payment Data Part D prescriber public use file: a methodological overview. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Prescriber_Methods.pdf. Accessed 9 Apr 2018.
44. CMS Office of Enterprise Data and Analytics. Medicare Fee-For-Service Provider Utilization & Payment Data Referring durable medical equipment, prosthetics, orthotics and supplies public use file: a methodological overview. https

://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/DME_Methodology.pdf. Accessed 9 Apr 2018.
45. CMS: Medicare Provider Utilization and Payment Data. Physician and other supplier. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html. Accessed 16 Aug 2016.
46. CMS: HCPCS—General Information. https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html. Accessed 13 Jan 2018.
47. CMS: Medicare Provider Utilization and Payment Data: Part D Prescriber. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html. Accessed 2 Dec 2017.
48. CMS: Medicare Provider Utilization and Payment Data. Referring durable medical equipment, prosthetics, orthotics and supplies. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/DME.html. Accessed 2 Dec 2017.
49. OIG: Office of Inspector General Exclusion Authorities US Department of Health and Human Services. https://oig.hhs.gov/. Accessed 1 Jan 2018.
50. OIG: Office of Inspector General Exclusion Authorities. https://oig.hhs.gov/exclusions/index.asp. Accessed 14 Jul 2016.
51. Pande V, Maas W. Physician medicare fraud: characteristics and consequences. Int J Pharm Healthcare Marke. 2013;7(1):8–33.
52. Bauder RA, Khoshgoftaar TM. A survey of medicare data processing and integration for fraud detection. In: 2018 IEEE 19th international conference on information reuse and integration (IRI). 2018. p. 9–14.
53. GPO: 31 U.S.C. 3729-FALSE CLAIMS. https://www.gpo.gov/fdsys/granule/USCODE-2011-title31/USCODE-2011-title31-subtitleIII-chap37-subchapIII-sec3729. Accessed 8 Jul 2018.
54. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. Mining data with rare events: a case study. In: 19th IEEE international conference on tools with artificial intelligence, 2007. ICTAI 2007, vol. 2. 2007. p. 132–9.
55. Khoshgoftaar TM, Seiffert C, Van Hulse J, Napolitano A, Folleco A. Learning with limited minority class data. In: Sixth international conference on machine learning and applications, 2007. ICMLA 2007. 2007. p. 348–3.
56. Van Hulse, J., Khoshgoftaar TM, Napolitano A. Experimental perspectives on learning from imbalanced data. In: Proceedings of the 24th international conference on machine learning. ACM; 2007. p. 935–42.
57. Apache: Linear Methods—RDD-based API. https://spark.apache.org/docs/latest/mllib-linear-methods.html. Accessed 10 Sept 2017.
58. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. Appl Stat. 1992;41:191–201.
59. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The weka data mining software: an update. ACM SIGKDD Explor Newslett. 2009;11(1):10–8.
60. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: practical machine learning tools and techniques. Burlington: Morgan Kaufmann; 2016.
61. le Cessie S, van Houwelingen JC. Ridge estimators in logistic regression. Appl Stat. 1992;41(1):191–201.
62. Apache Spark: classification and regression. https://spark.apache.org/docs/2.3.0/ml-classification-regression.html. Accessed 24 May 2018.
63. Apache Spark: Ensembles-RDD-based API. https://spark.apache.org/docs/2.3.0/mllib-ensembles.html. Accessed 24 May 2018.
64. Breiman L. Random forests. Machi Learn. 2001;45(1):5–32.
65. Bauder RA, Khoshgoftaar TM. Medicare fraud detection using random forest with class imbalanced big data. In: 2018 IEEE 19th international conference on information reuse and integration (IRI). 2018. p. 80–87.
66. Khoshgoftaar TM, Golawala M, Van Hulse J. An empirical study of learning from imbalanced data using random forest. In: ICTAI 2007. 19th IEEE international conference on tools with artificial intelligence, 2007, vol. 2. 2007. p. 310–7.
67. Bekkar M, Djemaa HK, Alitouche TA. Evaluation measures for models assessment over imbalanced datasets. J Inf Eng Appl. 2013;3:10.
68. Seliya N, Khoshgoftaar TM, Van Hulse J. A study on the relationships of classifier performance metrics. In: 21st international conference on tools with Artificial Intelligence, 2009. ICTAI'09. 2009. p. 59–66.
69. Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data–recommendations for the use of performance metrics. In: 2013 Humaine association conference on affective computing and intelligent interaction (ACII). 2013. p. 245–51.