A Novel Method for Fraudulent Medicare Claims Detection from Expected Payment Deviations

Richard A. Bauder, Taghi M. Khoshgoftaar
Florida Atlantic University
Email: {rbauder2014, khoshgof} @fau.edu

***Abstract*—**

Healthcare has and continues to be an integral component in people's lives, especially for the rising elderly population. One such healthcare program that provides for the needs of the elderly is Medicare. It is important that any such program be affordable but, unfortunately, this is not always the case. Out of the many possible factors for the rising cost of healthcare, fraud is a major contributor, but its impacts can be lessened through the use of fraud detection methods. We assess possible fraudulent activities by looking at the amounts paid to providers for services rendered to patients. In this study, we propose a novel methodology and framework towards identifying potential sources of fraud. We model these Medicare payments in order to create baseline values that reflect what the payments should be for a provider's specialty. We use these baseline expected payments and compare them to what was actually paid by Medicare for distinct specialties and healthcare services. Any deviations from the expected payments are flagged for further investigation. Our overall approach is consistent with related works, in healthcare, using anomaly-based detection methods to detect fraudulent activities, but we focus on an implementable and generalizable framework for initial fraud detection. Our results demonstrate the detection of possible fraudulent activities, with one specialty, Cardiology, demonstrating the detection of a known, real-world fraud case.

Keywords*—*Fraud Detection, Regression Analysis, Anomaly Detection, Medicare, Healthcare

I. Introduction

Healthcare is a critical component in most people's lives and as such, should be affordable. The need for healthcare is particularly important for the rising elderly population. They require increased healthcare and therefore, appropriate insurance coverage for various medical drugs and services. Medicare is one such insurance growth area for the elderly.

Medicare is a government program providing insurance to people over 65 years of age or certain younger individuals with specific medical conditions and disabilities [13]. Even with this obvious critical need, healthcare spending continues to rise with waste and abuse reduction efforts doing little to lessen these costs [12]. From the National Health Expenditures 2013 Highlights [11], released by the Centers for Medicare and Medicaid Services (CMS), US healthcare spending in 2013 increased 3.6% to reach $2.9 trillion. Medicare spending alone represented 20% of all national healthcare spending at about $587 billion, an increase of 3.4% from 2012. Furthermore, the Federal Bureau of Investigations (FBI) estimates that fraud accounts for 3-10% of all medical claims [24]. With the increases in healthcare costs, population growth, the inherent complexity of this program, and the huge volumes of money involved, this area has been, and continues to be, attractive for fraud and abuse activities.

The more recent public availability of Medicare and other Medicare-related datasets [4] asserts that the implementation of methods to detect fraudulent behaviors is not hindered by typical data privacy restrictions and access controls. The use of such large-scale data repositories and the smart application of data science (including data mining and machine learning activities) to detect fraud can lead to substantial cost recovery. For instance, it is estimated that with Medicare alone, recovery of 10% to 15% of expenses through fraud detection is possible [25]. A promising way to quickly assess possible fraudulent activities is by looking for events, cases, or values (e.g. payments, number of procedures) that diverge from an expected normal value, i.e. anomalous values. With this approach, a baseline is established which corresponds to expected values and is used to compare against actual values. This comparison can effectively illuminate abnormal values for further investigation in order to assess any true fraudulent behaviors. One key component of our general methodology is the efficient flagging of possible fraudulent activities, which can drastically reduce the amount of effort needed to locate these potentially aberrant values as well as hasten any subsequent investigations.

In this paper, we propose a methodology and framework to model average Medicare payment data in order to create expected payment values. Several models are created, out of which the best model is chosen, to generate these expected values for comparison with the corresponding actual values. The *Medicare Provider Utilization and Payment Data: Physician and Other Supplier Public Use File CY 2012 and 2013* (herein called "Physician and Other Supplier PUF") dataset is ingested and grouped by provider type (also called specialty), e.g. Cardiology, Healthcare Common Procedure Coding System (HCPCS) [3] code, and National Provider Identifier (NPI) [6]. We then train each model, per group, using 10-fold cross validation, repeated 4 times, on 90% of the data, with the remaining 10% used for validation. We discuss the results of both training and validation for five provider types to demonstrate the power and validity of this approach for the initial detection of possibly fraudulent activities, i.e. anomalous activities. Note the detection of these activities does not necessarily indicate fraud, but simply the possibility this event could be fraudulent.

978-1-5090-3207-5/16 $31.00 © 2016 IEEE
DOI 10.1109/IRI.2016.11

Our overall contribution is in creating a robust, implementable methodology and framework, as depicted in Figure 1, for the initial screening of fraudulent behaviors with detailed flagging and visualizations using Medicare data. Because our proposed methodology is focused on discovering possible fraudulent activities, there is no particular emphasis on any specific type of fraud [28] such as upcoding or self-referrals. Figure 1 illustrates the process necessary for handling the Medicare claims data, creating regression models to determine the expected payment values, assessing and flagging possibly fraudulent events (i.e. anomalies), and creating clear visualizations for further investigation into each event. The details of each process block in our framework are described further throughout the rest of this paper.

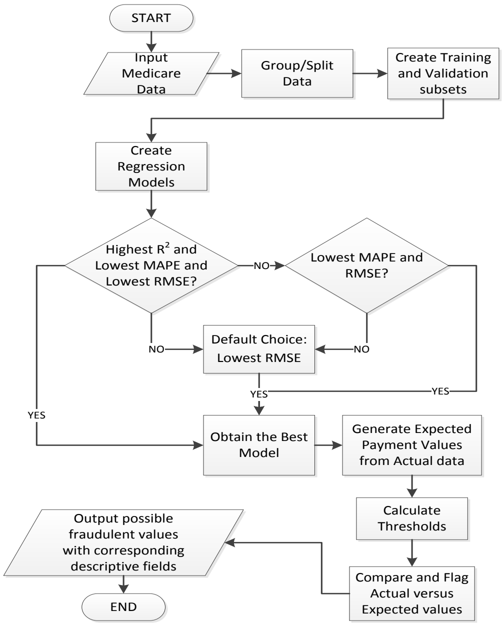
A key point, and something unique to our research, is using the large Medicare dataset of relatively long-term values to produce a reliable baseline model for expected payment predictions. Moreover, Thornton et al. [29] noted that most studies appear to focus on a specific area within healthcare, e.g. dentistry, but do not demonstrate the ability to generalize or scale beyond this single area. Our study applies general, easily scalable techniques on specific groups and data bins, across various healthcare areas, in order to reduce model variation, inaccurate expected payment values, and false positives. In testing our method, we successfully demonstrate that our method can detect potential fraud, comparing fraudulent activity flags against a known Cardiology fraud case in Florida [1], [5]. The real-world comparison indicates this initial screening could have been used to discover this fraudulent behavior earlier, reducing possible additional loss due to the fraud. To the best of our knowledge, only one prior study [29] has researched a version of expected versus actual comparisons for fraud detection on healthcare-related data. Two other studies [30], [20] have used an actual versus expected comparison method on either non-healthcare data or without a fraud detection focus.

The remainder of this paper is organized as follows. Section II discusses works related to the current research in this domain. In Section III, the general methodology used in this paper is detailed to include the dataset, models and performance metrics, as well as the framework design details. Section IV presents the results and discussions for our experiment. Finally, Section V outlines our conclusions and ideas for future work.

II. Related Work

Existing literature consists of numerous studies describing the analysis and detection of fraudulent activities in healthcare, primarily using private medical data records. There are few studies directly related to our work employing a baseline model and comparing these results to actual payment values. Furthermore, due to the recent, public release of the 2012 and 2013 Medicare datasets by CMS, research applied to these datasets remain nascent. There are several studies that specifically discuss expected versus actual value comparisons. In one study, Trnka [30] briefly discusses a method for detecting fraud using fictitious data in the area of agricultural development grants, applied to Six Sigma methodology [8]. In this case study, the author employs two different methods for detecting grant application anomalies. Anomaly detection, via the IBM SPSS Modeler, is used to generate anomaly indices in order to assess which records should be further investigated. Trnka also uses a neural network model to generate values for the expected farm's income. These expected values are then compared against actual income values in order to flag deviations for further exploration.

Fig. 1. Fraudulent Activity Detection Workflow


START
Input Medicare Data
Group/Split Data
Create Training and Validation subsets
Create Regression Models
Highest R^2 and Lowest MAPE and Lowest RMSE?
NO
Lowest MAPE and RMSE?
NO
Default Choice: Lowest RMSE
NO
YES
YES
YES
Obtain the Best Model
Generate Expected Payment Values from Actual data
Calculate Thresholds
Compare and Flag Actual versus Expected values
Output possible fraudulent values with corresponding descriptive fields
END

Thornton et al. [29] explores several outlier-based detection methods using Medicaid claims data, specifically for dental providers. The authors propose several methods to detect fraud using Medicaid data. In particular, their study involves multiple analysis techniques and outlier detection methods based on specific metrics, such as number of unique beneficiaries and claim payments, on a monthly basis. The authors propose several analysis techniques such as univariate and multivariate analysis, time-series analysis, and box plot analysis. The methods specific to outlier detection include the following: deviation from regression model, deviation clusters, single deviations from clusters, trend deviations, and peak deviations. The specific method most related to our current research, and the one we focus on, is the use and comparison of deviations from a regression model. They provide one example of a regression model with expected reimbursement values versus actual reimbursements with deviations, or outliers, indicated at 2.33 standard deviations away from the underlying regression model. It is not clear how they derived the 2.33 constant value. The authors provide a case study with 500 dental providers and claim the successful identification of 17 possibly fraudulent activities detected out of 360 records. Of these 17 possible fraudulent records, 12 are referred to officials for further investigations. The authors limit their method and experiment to a single specialty, dentistry, and do not offer any evidence or results for other specialties.

A study by Hu et al. [20] involves the application of both utilization profiling and anomaly detection. The authors use patient utilization data, specifically patients' clinical characteristics, to identify anomalous patterns. They generate expected patient utilization levels from observations using several regression models, then explain the difference between actual levels using Grubb's test. The models used to generate expected levels are Classification and Regression Trees, Random Forest, and Multivariate Adaptive Regression Splines. The use of Grubb's test is to assess the acceptable deviation ranges in order to find anomalous utilization levels. The authors demonstrate their method on 7,667 diabetes patients, detecting 51 anomalies. Their focus is on patient utilization, such as patient office visits, rather than provider claims. Hu et al. note their study is not focused on fraud, but they claim their anomaly method could be used as a potential indicator for fraud. The use of Grubb's test [10] to find acceptable utilization deviations assumes normality, which may not be applicable to their patient utilization dataset or other datasets, such as Medicare provider utilization. It is unclear as to whether the authors verified that their dataset had an approximately normal distribution, prior to using Grubb's test. Additionally, this test iterates over each value where multiple iterations can change the probabilities of detection. Thus, Grubb's test may not be the most robust method to detect anomalies using healthcare data.

Iyengar et al. [21] create a normalized baseline for the drug prescription focus area, in order to detect prescription drug fraud and abuse. The authors goal is to identify and rank possible audit targets from a database of prescription drug claims. They employ statistical hypothesis testing to identify entities that deviate from their expected behavior relative to the specific baseline model. To create their baseline model, the authors incorporate a rule list model structure and devise a rule generation algorithm for the model to detect prescription claims deviations. In order to focus their efforts, the authors are assisted by domain experts to cull relevant areas and features. The authors note their method is a preliminary approach to detecting abnormal behaviors for subsequent human investigations. They do not incorporate machine learning methods in their study, instead using rules created via a greedy search-like algorithm.

In our study, we, in part, build upon Trnka [30], Thornton et al. [29], and Hu et al. [20] in using an expected versus actual value construct, but focus on regression modeling using current Physician and Other Supplier PUF data. We differ from these related works in our methods for, and application of, data grouping across several specialties, best performing model selection, adaptive threshold estimates for deviations (with no inherent distribution assumptions), and visualizations. Moreover, we provide a general, implementable framework for the detection of fraudulent activities applying several techniques, in unique ways, across multiple provider specialties.

III. General Methodology

This section describes the Physician and Other Supplier PUF dataset, summarizes the regression models and performance metrics, and discusses the framework design. In particular, the latter section details the detection process as seen in Figure 1, further explaining our contribution.

A. Data

We use the Physician and Other Supplier PUF (2012 and 2013) dataset provided by the Centers for Medicare and Medicaid Services [4]. Two primary features for identification of providers and procedures performed are Healthcare Common Procedure Coding System (HCPCS) codes and National Provider Identifier (NPI) [6] numbers. Healthcare Common Procedure Coding System codes are used to identify a provider's specific medical service. For example, the range of codes from R0070 to R0076 are used for diagnostic radiology services. The NPI is a unique 10-digit identification number for healthcare providers issued by the Centers for Medicare and Medicaid Services. For privacy reasons, the NPI numbers are purposefully obfuscated in this paper. To provide a better understanding of this dataset, the general process to create and pay a claim is briefly described. After a patient is admitted, the conditions are assessed and the services are provided. The healthcare provider staff (physicians in most cases) annotate this on the patient's medical chart. A medical coder takes the information on these charts and translates them into the appropriate diagnoses and procedures (coded as HCPCS) in order to create and file a claim. Based on this information, including the HCPCS codes, Medicare processes the claim and makes the appropriate payment.

The entire dataset represents 6,452 distinct procedure codes performed by 968,589 physicians in the United States, with the dataset having a total of 24,406,332 instances and 29 features. In order to construct our initial fraud detection method from this long-term data, we filter the Medicare dataset for non-facility (typically considered an office environment) and non-prescription data in Florida. The non-prescription data are those HCPCS codes that are not for specific services listed on the Medicare Part B Drug Average Sales Price file [2], thus are actual provider services versus drug-specific activities and/or prescriptions. This Florida only subset consists of 2,766 procedures codes, 44,077 physicians and 83 provider types, with 795,241 instances. From the 29 total features, 13 features are used for our expected payment prediction and fraudulent activity detection method. Out of these 13 features, 6 are used for the regression models, while the remaining features are used for the identification of possibly fraudulent activities and corresponding provider(s). An ANalysis of VAriance (ANOVA) test was performed for each group. An ANOVA is essentially a general t-test applied to more than two groups to test the null hypothesis for variable (or group) significance. All features chosen for the regression models are statistically significant, except for *Year*, which is not considered significant for Vascular Surgery, Thoracic Surgery, and Family Practice. We suspect adding additional years and/or other time-based values will increase variable significance and better model and detect temporal trends. At this time, we are limited to two years of available Medicare data. Table I lists and describes these features, both used in the model and for descriptive purposes, with an (*) indicating the feature used in the regression models.

Based on the information provided by CMS [4], this dataset contains line items for physician and supplier Part B fee-for-service claims only. It does not include claims for Durable Medical Equipment, Prosthetics, or Orthotics. Additionally, for this study, we did not incorporate other Medicare and Medicare-related datasets that could include information such

TABLE I. DESCRIPTION OF MEDICARE FEATURES

Feature	Description
Provider Type	Medical provider's specialty, e.g. Cardiology
HCPCS Code	Code for specific medical service furnished by the provider
NPI	Unique provider identification number
Last Name	Provider's last name
Address	Provider's office address
City	Provider's office city
Zip Code*	Provider's office zip code
Year*	Year procedures were performed (2012 or 2013)
Line Service Count*	The number of services provided
Beneficiary Day Service Count*	The number of distinct Medicare beneficiary/per day services
Avg. Medicare Allowed Amount*	Medicare payments, deductible/coinsurance, and third party amounts
Avg. Medicare Payment Amount*	Amount Medicare paid the provider for services performed

as physician referrals or prescriptions. These additional data are considered for future work to detect additional fraudulent behaviors.

B. Models and Performance Metrics

We selected five different regression models due to their varied underlying model logic, general applicability, and prior experience. Even though these five are used, we are not limited to only these model choices. This paper is not a study on regression models nor do we propose model improvements or enhancements, but rather a discussion of the framework. Thus, we only briefly describe them in this section. The interested reader can find general information on regression models and analysis in [15], [22], [31], [23], as well as specific references for each of the models in this section.

The Support Vector Machine (SVM) was originally created for classification using so-called support vectors to maximally separate data into distinct hyperplanes, for each category. SVM performs linear classification, but, with the use of a kernel trick, can be used for non-linear classification. An extension of SVM classification for regression is called Support Vector Regression (SVR) [27]. The model produced by classification depends only on a subset of the training data, as defined by each support vector, thus data points beyond this are not considered for classification. Similarly, SVR ignores most of the training data already close to the model prediction. An SVM prediction entails every variable being multiplied by the corresponding element of every support vector. This model is referred to as SVM or linearSVM in this study.

A Generalized Linear Model (GLM) is a linear model using link functions [14]. These link functions are used to linearly relate the dependent variables to predictor, or independent, variables. An example of a link function is the *logit* function which acts as a binary response variable for ordinal values. GLMs are multiple regression models quantifying relationships between multiple independent and/or dependent variables. A variation on the standard GLM is a model fit with lasso or elasticnet regularization via penalized maximum likelihood [16], referred to as GLMNET. This model is more robust than a traditional GLM, capable of efficiently handling large, sparse matrices. The final variation on the traditional GLM incorporates Bayesian inference with independent normal or Cauchy prior distributions for model coefficients [19]. This proper prior distribution produces stable, regularized estimates capable of performing well on real-world applied work. In our paper, this method is called BayesGLM.

Multivariate Adaptive Regression Splines (MARS) is a non-parametric regression model that accounts for the nonlinearities between variables and their interactions [17]. MARS utilizes a hinge function as piecewise linear functions (fitting the data), as well as non-linear functions for variable relationships through combined hinge functions. MARS performs automatic variable selection, is suitable for large datasets, and is more flexible than traditional linear models.

Performance metrics are the same across the aforementioned regression models to effectively compare model performance in order to select the best model per provider type group. These metrics are standard in regression analysis, so we only briefly describe the metrics used in our study in Table II.

TABLE II. REGRESSION PERFORMANCE METRICS

Metric	Description
Pearson Correlation Coefficient (r)	Linear relationship between pairs of variables
Coefficient of Determination (R^2)	The variance in the dependent variable explained by the independent variable
Root Mean Squared Error (RMSE)	Estimated standard deviation of unexplainable variations in the dependent variable
Mean Absolute Percentage Error (MAPE)	Average percentage difference between actual and observed values
Mean Absolute Error (MAE)	Average of the absolute differences of predicted and observed values

C. Framework Design

We designed and implemented our methodology and framework, as seen in Figure 1, and experiments using the R language [26]. In order to generate the most meaningful and accurate expected payment values, we split our Florida-only dataset into distinct groups by provider types (or specialties) and HCPCS codes. Each of these groups is then further split into training and validation data. The Classification And REgression Training (CARET) [18] package is used to split the data as well as provide an extensible framework for model creation. CARET is a set of functions to streamline the process for creating predictive models. Furthermore, for this study, we use the default model parameters, thus do not consider model tuning. This is not the focus of the work and is an option for future work.

We limit the experiments in this initial study to 6 of the 83 provider types, which include Thoracic Surgery, Family Practice, Cardiology, Vascular Surgery, Hand Surgery, and Pediatric Medicine. In choosing the provider types for our experiments, we sampled based on characteristics such as high average Medicare payments, a low number of instances, high number of procedures performed, and high and low number of procedure codes used. Our aim was to use a range of diverse provider types within the Florida Medicare dataset. Table III lists the provider types with the corresponding selection characteristics.

As previously mentioned, the five regression models are created from the 90% training data for each provider type, e.g. Cardiology. Note the purpose of splitting the data, within each group, and training on 90% of the data is to a) create the baseline models that produce the expected payment values,

TABLE III. Summary of Provider Types and Characteristics

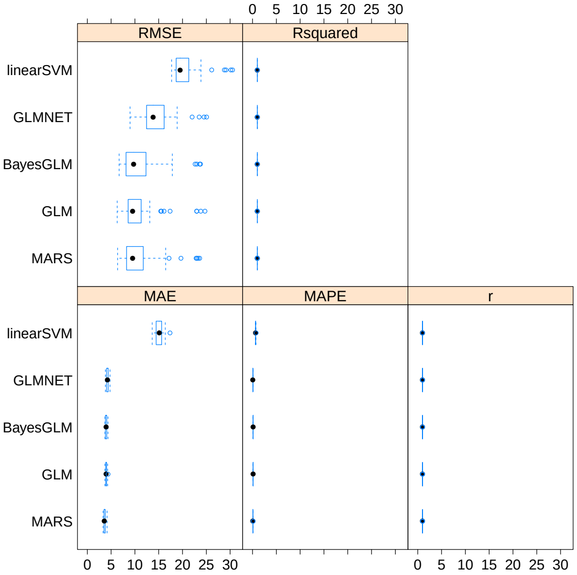
Provider Type	Selection Characteristic
Thoracic Surgery	High average Medicare payment and low number of instances
Vascular Surgery	High average Medicare payment and low number of instances
Family Practice	Large number of different procedures performed and Diagnosis-Related Group
Cardiology	Average number of different procedures performed and high number of instances
Hand Surgery	Low quantities of procedures performed and low number of instances
Pediatric Medicine	Low quantities of procedures performed and low number of instances

and b) generate cross-validation metrics for the selection of the best baseline model. With that, validation is done using the remaining 10% of the data as a proxy for new, actual data when comparing against the model-generated expected payment values. The performance results are generated via 10-fold cross validation repeated 4 times and averaged to get the final model metrics. A sample visualization for a model comparisons using the metrics can be seen for the Cardiology provider type in Figure 2.

All trained models produce these metrics for comparison. More specifically, a pairwise comparison is performed to choose the best performing model. These comparisons are made hierarchically based on the following conditions:

```
if highest $R^2$ and lowest RMSE and lowest MAPE then
    return Best Model
else
    if lowest RMSE and lowest MAPE then
        return Best Model
    else
        if lowest RMSE then
            return Best Model (default)
        end if
    end if
end if
```

Fig. 2. Training Model Metrics for Cardiology



Following the selection of the best performing model, the remaining 10% of the dataset is used for validation. This validation step mimics the comparison to new, real-world data in lieu of updated Medicare data, i.e. Medicare data covering 2014 or 2015 payments. To create the expected payment values per group, the actual payment values from the validation subset are passed into the best performing model, which produces the corresponding expected values. The expected value is what we assume to be the correct payment amount for a specific provider type and procedure. We now have a vector of expected payments (from the model) and another of comparable actual payments (from the validation subset). The residuals, or errors, from this model are used to created the thresholds from which to assess deviations for possible fraudulent activities. Each group is done in a standard way, yet encompasses the unique distribution for a given provider type. The provider type's data uniqueness is further exploited by binning the data by the number of procedures performed, since this number could represent a different payment distribution. For example, a lot of procedures performed could indicate a higher average Medicare payment, but certain provider types and/or procedure codes could have high average payments regardless of the number of procedures performed. The binning by procedures performed account for some of these variations and sets the thresholds adaptively. The binning by the count of procedures performed is presently dependent on the Medicare data, with future research working to generalize this process. Even though the process for creating thresholds with Medicare data is adaptive, there may still be a need for a user-defined adjustment factor. This factor could be used for specific provider types that have deviations beyond those fit by the long-term regression model, or unique circumstances that require a greater degree of variation versus past payments. It is important in any such framework to allow for some user feedback and adjustments to quickly adapt to any critical changes. The process for creating thresholds for a provider type is detailed in the following steps:

1) Calculate the errors $\epsilon_{group} = \hat{y} - y$
2) Compute the average error μ_{group}
3) Calculate the standard deviation of the errors σ_{group}
4) Create a user-defined adjustment factor, x, with a default value of 2
5) Bin data by the number of procedures performed, i
    a) Bin sizes are calculated by cutting the total number of procedures by $\lfloor median(i) \rfloor$
6) Calculate the threshold value per bin by:
    a) Get the standard deviation of errors σ_{bin}
    b) The threshold is calculated as $\gamma_{bin} = \sigma_{group} + (x \times \sigma_{bin})$
    c) Impute an average threshold value for any missing values

Once the thresholds are created for each bin, a flag is set indicating deviations beyond these thresholds thus possibly fraudulent activities (or anomalous events that do not necessarily indicate an actual fraud event). The flags are set to "good" by default (indicating no fraudulent activities detected based on the threshold deviations), where payment values of $|\epsilon_{bin}| > \gamma_{bin}$ are flagged as possible fraud. Using fraud visualizations, these thresholds are depicted by vertical segments around the expected payment values, similar to error bars. Therefore, any actual payment values that are outside of these thresholds will be seen as possibly fraudulent. For this study, we are only interested in actual values that are greater than the upper threshold of the expected payment values. If the actual value is above the threshold, this could indicate that the Medicare payment made was more than what was expected (per the baseline model), thus possibly fraudulent in nature. Actual payment values below the lower threshold of the expected payments could show another form of fraud or misuse behavior, such as undercoding [9]. The flagged fraudulent activity values can then be associated with various characteristics such as NPI, last name, address, and zip code. These are readily available identification metrics for these fraudulent activities that can be used to either ignore this particular flag or investigate further.

Finally, similar to Thornton et al. [29], we consider a rate of detection for possible fraudulent activities, as the number of flagged values over the total number of values, per grouping. Future work would include additional comparisons to real-world cases and/or subject matter expert evaluations in order to accurately estimate a corresponding success rate, i.e. the number of actual fraud cases detected.

IV. Results and Discussion

The results for our fraud detection framework, across the six chosen provider types, indicate our method can efficiently and effectively detect deviations from expected payments. The flagging of possible fraudulent activities in the actual data, that show sufficient deviation from the baseline expected Medicare payments, can substantially reduce the effort needed to manually review claims. Tables IV and V present the training and validation summaries and validation results, respectively. The summaries show possible patterns relating the underlying data distributions (Medicare payments and procedure counts) to the selected regression model. The majority of the chosen models are MARS, which is reasonable given its flexibility over traditional linear models, but two other provider types did not choose MARS instead selecting GLM. This asserts the need to utilize general modeling techniques in specific ways, for specific data, which has not been discussed in detail in any related works. In Table V, the possible fraud cases flagged (what we call flag rate), appears to be below 1.50%, for most specialties, meaning that each provider type has a limited number of possible fraudulent activities to investigate. This reduction from the original data size allows for more focused investigations on the top priority cases, e.g. more than a 99% reduction in the case of Cardiology. Out of the six provider types, Family Practice differs most from the others primarily in the field's generality and the number of different procedures performed. This is reflected in the number of flagged fraud cases and flag rate versus the other provider types, which are more specialized in function. The less homogeneous nature of Family Practice procedure codes can also be seen in the large threshold ranges (the large vertical lines around the expected payment value) in Figure 5.

The primary output of our framework is flagged fraudulent behavior, but other interim views can also be used to add more detail and possible value for fraud detection. Figure 3 depicts an interim view for Cardiology that is divided by group as either "good" or possible fraud. This shows the distribution of fraudulent activity values, to include payment and procedure count density and ranges. Additional focus areas can be assessed based on where most of the fraudulent flags occur for model adjustments and/or further investigations.

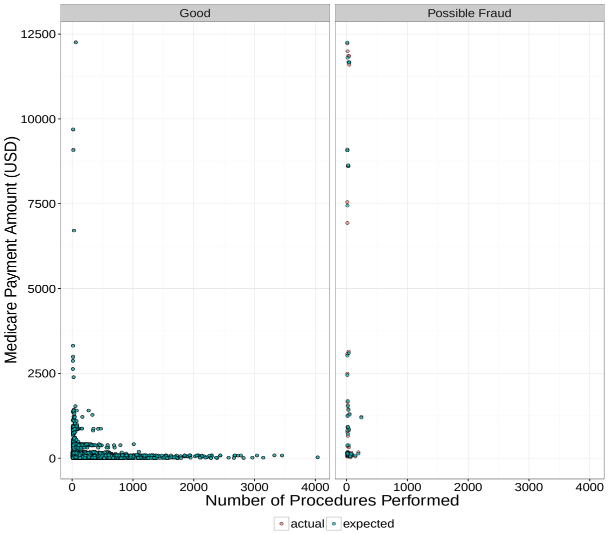
TABLE IV. Training and Validation Model Summary

Provider Type	Instances (Training)	Instances (Validation)	Chosen Model
Thoracic Surgery	685	76	GLM
Vascular Surgery	3,717	413	GLM
Family Practice	81,513	9,057	MARS
Cardiology	46,242	5,137	MARS
Hand Surgery	2,113	234	MARS
Pediatric Medicine	772	85	MARS

TABLE V. Fraud Validation Results by Provider Type

Provider Type	# of Possible Fraud Cases	Flag Rate
Thoracic Surgery	1	1.32%
Vascular Surgery	3	1.43%
Family Practice	229	2.53%
Cardiology	22	0.43%
Hand Surgery	3	1.28%
Pediatric Medicine	1	1.18%

Fig. 3. Cardiology - Possible Fraud Facet Plot



Figures 4, 5, and 6 illustrate the specific, flagged fraudulent activities with identification labels for Vascular Surgery, Family Practice, and Cardiology. For clarity, only a subset of the flagged values are labeled with these identification numbers, which are placeholders for masked NPI numbers. Additionally, none of the other descriptive features are currently used for these visualizations. Thoracic Surgery and Pediatric Medicine produced only one flagged value each, thus do not require visualizations. Hand Surgery did not have any flagged values greater than the upper threshold, but there were three deviations below the lower threshold value.

The Figures illustrate the power of the binning process, within each provider type group, in producing the thresholds and adapting to the specific data distribution per provider type.

Fig. 4. Vascular Surgery - Fraudulent Activity Flagged Values

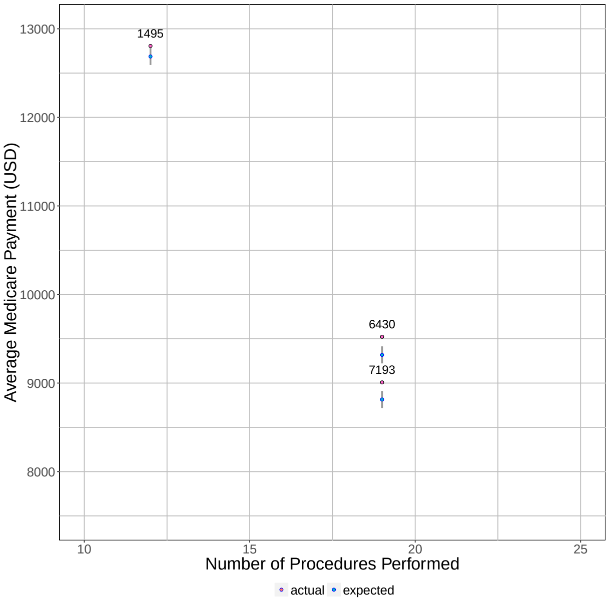
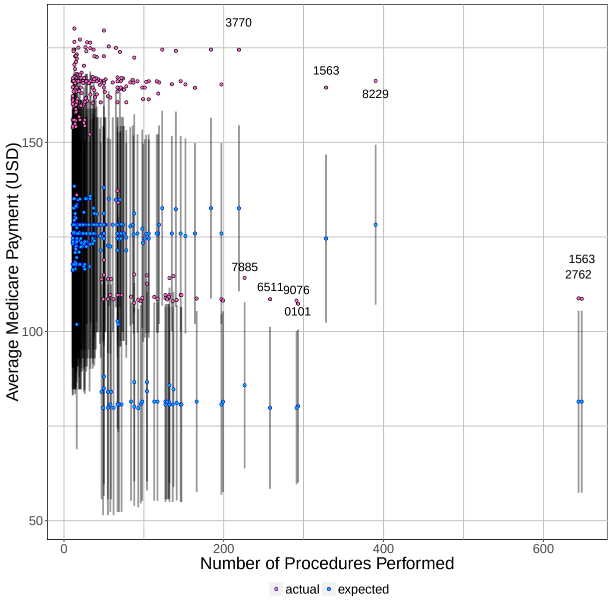
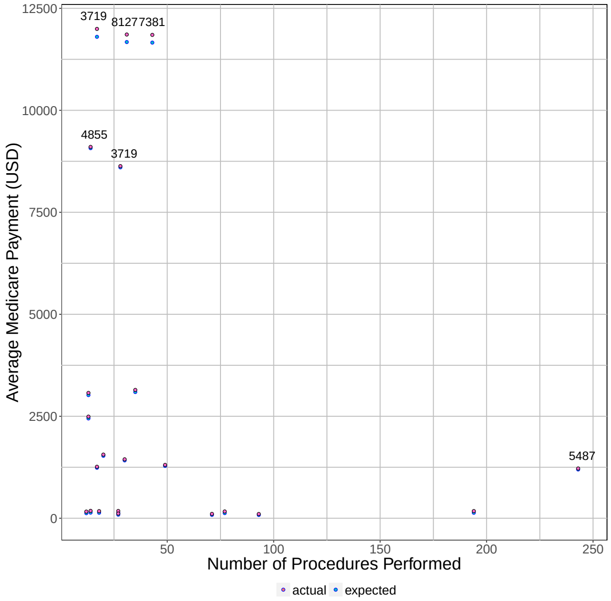


Fig. 5. Family Practice - Fraudulent Activity Flagged Values



This can be seen clearly in Figure 5 by how the deviations are grouped in the plot both horizontally and vertically, as well as changes to the thresholds as indicated by the lengths of the gray vertical lines. Therefore, this binning process helps narrow the detection focus by adjusting the flagged deviations and what payments are deemed acceptable. Across these provider types, the vertical bar sizes show the differences in thresholds generated and thus the uniqueness that needs to be captured by any model to effectively detect fraudulent behaviors.

Fig. 6. Cardiology - Fraudulent Activity Flagged Values



To further support our methodology and framework, based on the experiment results, we use the NPI numbers for the flagged values and search for known fraud cases in news reports and the U.S. Department of Health & Human Service Office of Inspector General Exclusion List [7]. When searching for providers via last name and NPI, we discovered a Cardiologist under investigation for fraud [1], [5]. It was reported that this provider billed Medicare for medically unnecessary peripheral artery interventions, which is flagged by our method, with a lable of 5487, as seen in Figure 6. These unnecessary billings would show up as deviations from expected payment values for a particular provider type given enough data to create the baseline model. We include two years of Medicare data to create these models and create stable, reliable expected payment predictions. This one observed example of a real-world, reported fraud case demonstrates the potential of our framework for early detection and investigation of fraudulent activities. Even given this successful detection, due to the limited number of real-world fraud cases pertaining to the current 2012 - 2013 Medicare claims data, there is a gap in assessing our framework's comparative performance with relation to known cases. Additionally, the publicly available Medicare claims data does not provide labels indicating known fraudulent activities. Therefore, our framework is a robust way to explore possible fraudulent events in order to provide a solid basis for further investigation.

V. Conclusion

Medicare fraud continues to be a burden on the U.S. healthcare systems requiring novel and practical solutions. There are many studies demonstrating possible ways of detecting fraudulent behaviors in healthcare, but very few focusing on Medicare. The recent public availability of Medicare data is a boon for continued research into methods to detect, investigate, and reduce healthcare fraud. In this study, we create a novel methodology and framework for an implementable fraudulent activity detection system. Our method uses well-known, general regression methods (e.g. Support Vector Machines) and analyses on distinct data groups (such as Cardiology and Family Practice), to produce results indicating possible fraud by detecting abnormal payment behaviors. More specifically, we employ several techniques to improve the flag rate of fraudulent behaviors to include automatic model selection, data grouping and binning, and adaptive thresholds. Additionally, the use of a reasonably long-term data (about two years) produces stable baseline models for expected payment value predictions. Our framework, as seen in Figure 1, is effective, yet simple enough to be implemented and used in real-world, fraud monitoring situations.

Our experiments indicate a flag rate between 0.43% and 2.53% on a range of varied provider types. The model selection process shows some provider types, due to their underlying payment and procedure distributions, require different models. The experiments produced four uses of MARS and two of GLM. Furthermore, the use of adaptive thresholds through grouping and binning, for each provider type, further refine the resulting fraudulent activity flags, taking into account the differences in both payments and number of procedures unique to provider types. The detection of a real-world fraud case further supports the validity of our methodology in its ability to detect potential fraud.

The methodology presented demonstrates that Medicare fraud detection is possible with a noted real-world detected fraud case. Even so, ongoing research and future work will involve refinements to the provider type grouping and binning, as well as differing threshold methods, such as using distance measures. Furthermore, we intend to generalize the binning process beyond only Medicare claims data. Additionally, integrating additional datasets, such as referrals, could be assessed in order to detect a wider range of fraudulent activities. Finally, as previously mentioned, additional validation and testing against real-world cases should be performed to verify and enhance our current detection methodology, as well as expanding the Medicare dataset beyond Florida and incorporating any newly released Medicare data.

Acknowledgment

We acknowledge partial support by the NSF (CNS-1427536). Opinions, findings, conclusions, or recommendations in this material are the authors' and do not reflect the views of the NSF.

References

[1] "Cardiologist plagued by legal woes files for Chapter 11 bankruptcy protection." [Online]. Available: http://www.ocala.com/article/20160422/ARTICLES/160429933

[2] Centers for Medicare and Medicaid Services Frequenty Asked Questions. [Online]. Available: https://questions.cms.gov/

[3] Centers for Medicare and Medicaid Services: HCPCS General Information. [Online]. Available: https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/

[4] Centers for Medicare and Medicaid Services: Research, Statistics, Data, and Systems. [Online]. Available: https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems.html

[5] Government Intervenes in Lawsuit Against Florida Cardiologist Alleging Unnecessary Peripheral Artery Interventions and Payment of Kickbacks. [Online]. Available: https://www.justice.gov/opa/pr/government-intervenes-lawsuit-against-florida-cardiologist-alleging-unnecessary-peripheral/

[6] National Plan & Provider Enumeration System (NPPES): National Provider Identifier. [Online]. Available: https://nppes.cms.hhs.gov/NPPES/

[7] Office of Inspector General: Exclusions Program. [Online]. Available: http://oig.hhs.gov/exclusions/index.asp

[8] "What Is Six Sigma?" [Online]. Available: https://www.isixsigma.com/new-to-six-sigma/getting-started/what-six-sigma/

[9] "Common Problems in Medical Coding," 2013. [Online]. Available: http://www.medicalbillingandcoding.org/common-problems-coding/

[10] "NIST/SEMATECH e-Handbook of Statistical Methods," 2013. [Online]. Available: http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

[11] "National Health Accounts by service type and funding source," 2014. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-systems/Statistics-Trends-and-reports/NationalHealthExpendData/index.html

[12] "The facts about rising health care costs," 2015. [Online]. Available: http://www.aetna.com/health-reform-connection/aetnas-vision/facts-about-costs.html

[13] "US Medicare Program," 2016. [Online]. Available: https://www.medicare.gov

[14] A. J. Dobson and A. Barnett, *An introduction to generalized linear models*. CRC press, 2008.

[15] J. Fox, *Applied regression analysis and generalized linear models*. Sage Publications, 2015.

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[17] J. H. Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, pp. 1–67, 1991.

[18] M. K. C. from Jed Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, and C. Candan, *caret: Classification and Regression Training*, 2016, R package version 6.0-68. [Online]. Available: https://CRAN.R-project.org/package=caret

[19] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, pp. 1360–1383, 2008.

[20] J. Hu, F. Wang, J. Sun, R. Sorrentino, and S. Ebadollahi, "A healthcare utilization analysis framework for hot spotting and contextual anomaly detection." in *AMIA*, 2012.

[21] V. S. Iyengar, K. B. Hermiz, and R. Natarajan, "Computer-aided auditing of prescription drug claims," *Health care management science*, vol. 17, no. 3, pp. 203–214, 2014.

[22] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[23] T. M. Khoshgoftaar and N. Seliya, "Fault prediction modeling for software quality estimation: Comparing commonly used techniques," *Empirical Software Engineering*, vol. 8, no. 3, pp. 255–283, 2003.

[24] L. Morris, "Combating Fraud In Health Care: An Essential Component Of Any Cost Containment Strategy," 2009. [Online]. Available: http://content.healthaffairs.org/content/28/5/1351.full

[25] D. Munro, "Annual U.S. healthcare spending hits $3.8 trillion," 2014. [Online]. Available: http://www.forbes.com/sites/danmunro/2014/02/02/annual-u-s-healthcare-spending-hits-3-8-trillion/

[26] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: https://www.R-project.org/

[27] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[28] T. Swanson, "The 5 most common types of medical billing fraud," 2012. [Online]. Available: http://www.business2community.com/health-wellness/the-5-most-common-types-of-medical-billing-fraud-0234197

[29] D. Thornton, G. Capelleveen, M. Poel, J. Hillegersberg, and R. M. Müller, “Outlier-based health insurance fraud detection for us medicaid data,” 2014.

[30] A. Trnka, “Six sigma methodology with fraud detection,” in *9th WSEAS Interanational Conference on Data Networks, Communications, Computers (DNCOCO10): University of Algarve, Faro, Portugal*, 2010, pp. 162–165.

[31] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.