Health Serv Outcomes Res Method (2017) 17:256–289
DOI 10.1007/s10742-017-0172-1

Multivariate outlier detection in medicare claims payments applying probabilistic programming methods

Richard A. Bauder[1] · **Taghi M. Khoshgoftaar**[1]

Received: 5 October 2016 / Revised: 29 May 2017 / Accepted: 12 June 2017 /
Published online: 20 June 2017

Abstract The rising elderly population continues to demand more cost-effective healthcare programs. In particular, Medicare is a vital program serving the needs of the elderly in the United States. The growing number of people enrolled in healthcare programs such as Medicare, along with the enormous volume of money in the healthcare industry, increases the appeal for, and risk of, fraudulent activities. 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. In this paper, we assess possible illegitimate activities by looking at the amounts paid to providers for services rendered to patients. We propose a novel method for fraud detection that focuses on discovering outliers in Medicare payment data using multiple predictors as model inputs. Our multivariate outlier detection approach is twofold: (1) create a Multivariate Adaptive Regression Splines model to produce studentized residuals and, (2) use the residuals as input into a general univariate outlier detection model, based on full Bayesian inference, using probabilistic programming. Using this approach, we are able to incorporate multiple variables to detect outliers with a model that provides probability distributions, with credible intervals, rather than just point values, as with most common outlier detection methods. Additionally, these credible intervals further enhance confidence that the detected outliers should in fact be considered outlying values, thus possibly fraudulent activities. Our results show that the successful detection of these possibly fraudulent activities can provide effective and meaningful results for further investigation within various medical specialties.

Keywords Medicare fraud · Multivariate outlier detection · Regression · Bayesian inference · Probabilistic programming

✉ Richard A. Bauder
rbauder2014@fau.edu

Taghi M. Khoshgoftaar
khoshgof@fau.edu

[1] Florida Atlantic University, Boca Raton, FL, USA

1 Introduction

Healthcare is a crucial component in our daily lives, yet its importance is further exacerbated in the lives of the rising elderly population. In general, they require increased healthcare and therefore, appropriate insurance coverage for various medical drugs and services[1] (Matthews 2013). It is critical for these healthcare insurance programs to be affordable, but unfortunately, this is not always the case. The increasing costs associated with most healthcare plans can financially cripple individuals and families. Even given such social prominence, healthcare costs continue to rise with fraud, waste, and abuse (FWA) reduction efforts doing little to diminish these costs.[2]

Medicare is a US government program providing for the insurance needs of people over the age of 65, or younger individuals with certain medical conditions and disabilities.[3] The population of individuals age 65 years or older represented 14.5% of the US population in 2014. The number of elderly individuals rose 28% since 2004 versus just 6.5% of those under 65 years of age.[4] To provide some context in US dollars, the National Health Expenditures 2013 Highlights,[5] released by the Centers for Medicare and Medicaid (CMS),[6] indicate the US healthcare spending in 2013 increased 3.6%, from 2012, 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 (Morris 2009), with similar figures reflected in a study by Berwick and Hackbarth (2012). CMS states no official fraud estimates, relying on the FBI figures (Greenburg 2013). With the increases in healthcare costs, population growth, and the large quantities of money involved, this area has been, and continues to be, attractive for fraud and abuse.

The public availability of Medicare and other Medicare-related datasets[7] allows for the implementation of methods to detect improper behaviors that are not hindered by typical data privacy restrictions and access controls, as with most patient health records. 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 10–15% of expenses through fraud detection is possible (Munro 2014). Even so, the CMS has no reliable way to measure the impacts of fraud, though there are several methods that can help reduce potentially fraudulent activities (Steinbusch et al. 2007). Another important factor contributing to the difficulty in detecting fraud is the lack of current and available fraud-related labeled records. Most of the labeled records come from insurance carriers and, as such, are limited due to security and privacy concerns, as well as corporate interests. Beyond these privacy concerns, in order to create labeled records for fraudulent activities, a subject matter expert audits medical data on a limited number of claims, due to the large volume of claims. These limitations elicit a need to look for anomalies in

[1] http://www.pbs.org/newshour/rundown/how-growth-of-elderly-population-in-us-compares-with-other-countries/.

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

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

[4] http://www.aoa.acl.gov/Aging_Statistics/Profile/2015/.

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

[6] https://questions.cms.gov/.

[7] https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems.html.

healthcare data to detect and flag possible fraudulent events. Looking for anomalies in the data greatly reduces the need for labeled records, while still providing information for flagging outliers indicating potential FWA activities.

In this paper, we propose an innovative multivariate outlier detection method combining multivariate regression and probabilistic programming, with a novel application of this twofold approach. We employ a Multivariate Adaptive Regression Splines (MARS) model to fit multiple predictor (independent) variables with a single response (dependent) variable. The MARS model residuals are used as inputs to our generalizable, fully Bayesian probability model to detect anomalous values (or outliers). In general, probabilistic programming allows for the creation of a probability model to fully represent uncertainty, or variability, about any underlying information explaining observed data, for probabilistic inference. Additionally, these models enable the incorporation of prior knowledge and/or assumptions, such as from a physician, and probability distributions for each data point in assessing potential outliers. In our study, we are interested in Medicare fraud thus consider Medicare claims payment data as the response variable. With our proposed method, the resulting probability distributions, per claim, provide more information and give greater confidence in flagging illegitimate claims. It is important to reiterate that the emphasis should be on the need for further investigation into each outlier. The identification of fraudulent activities, via outlier detection, still requires human involvement. Even so, outlier detection does reduce the number of possible observations that will require additional inspection and corroboration, thus allowing for better utilization of time and resources. In this paper, we do not discuss different types of healthcare fraud (Swanson 2012), such as upcoding or self-referrals, but the reader is referred to Bauder et al. (2017) for additional information.

In order to demonstrate the efficacy of our method, we provide two comparative analyses. For both, we use real-world healthcare data from the *Medicare Provider Utilization and Payment Data: Physician and Other Supplier Public Use File CY 2012, 2013, and 2014* (herein called "Physician and Other Supplier PUF" or "Medicare data").[8] The first analysis demonstrates our probability model and compares it to other common univariate outlier detection methods. We show that our model not only detects possible outliers, but more importantly, provides a distribution of probabilities, with credible intervals, for each value with which to determine an outlier. This probability distribution is different than the point values returned by most of the other outlier detection methods. The distributions indicate the likelihood of a value being an outlier, as well as the confidence in that outlier's probability result. The second analysis focuses on multivariate outlier detection using our method and comparing it against methods incorporating Mahalanobis distance and k-means clustering, both popular ways to detect outliers with multiple variables. Our results indicate our multivariate outlier detection method favorably compares to the other methods, while providing more meaningful information on each potential outlier. We follow the comparative analyses with an application of our method using Medicare data. To the best of our knowledge, there are no other studies that use both regression and probabilistic programming models for multivariate outlier detection in Medicare (or healthcare) data.

The rest of the paper is organized as follows. Section 2 discusses works related to the current research. In Sect. 3, we review several current univariate and multivariate outlier detection methods. In Sect. 4, we examine our proposed outlier detection approach, combining both a MARS regression model and our probability model. Section 5 describes

[8] https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems.html.

the Medicare data used in our study. Section 6 reviews the multivariate outlier detection method and experimental design. Section 7 discusses the univariate and multivariate comparative analyses. In Sect. 8, we provide further discussions pertaining to the application of our model using Medicare data. Finally, Sect. 9 outlines our conclusions and ideas for future work.

2 Related works

There have been numerous studies on the effects of outliers on regression models, focused primarily on how to detect and handle these outlying values. A commonality seen throughout a majority of these papers (Cousineau and Chartier 2015; Stevens 1984; Nedret and Gulsen 2008; Das and Gogoi 2016) is the use of outlier detection methods to help improve linear regression models. We are not concerned with how outliers affect the regression model, but rather how to detect outliers based on the output of a regression model. Techniques such as studentized residuals and Mahalanobis distance, can be used to assess possible outliers in the output of a regression model. The use of several common techniques, such as studentized residuals, to detect outliers in order to improve regression models, is discussed in a paper by Mínguez et al. (2012). The authors employ detection methods to automatically find outliers, such as hurricanes, that may be present in instrumental records (e.g. buoys), to protect the results of an analysis from these rare events (i.e. either remove or note these events to create a clean, baseline dataset for analysis). Chaloner and Brant (1988) present a Bayesian approach using residuals in regression models to detect outliers. The authors assess outliers using the standardized residuals, from a simple linear regression model, and assumed non-informative priors using the posterior distribution for probabilities of outliers. These posteriors are modeled as Student's *t* distributions to better capture outlying values towards the tails of the distribution. The results show the Bayesian approach provides similar results to other methods, such as standardized residuals and predictive discordancy diagnostics, thus it is difficult to assess the value of these methods.

The existing literature incorporating statistical or machine learning techniques to detect outliers in healthcare is fairly extensive, using common methods such as decision trees or neural networks (Li et al. 2008; Joudaki et al. 2014; Tomar and Agarwal 2013). We do not detail the aforesaid statistical and machine learning techniques, but the interested reader can obtain more information in the referenced survey papers and in Witten and Frank (2005). As with our work, several papers use distributions of expected versus actual values (i.e. model residuals) to assess various types of possibly fraudulent behaviors. In one study, Trnka (2010) briefly discusses a method for detecting fraud using fictitious data in the area of agricultural development grants, applied to Six Sigma methodology (Pande and Neuman 2000). First, the author employs anomaly detection, via the IBM SPSS Modeler, to generate anomaly indices to assess which records should be further investigated. Trnka then uses a neural network model to generate values for the farm's expected income, based on multiple input variables. These expected values are then compared against actual income values in order to flag deviations for further exploration.

Thornton et al. (2014) explores several outlier-based detection frameworks using Medicaid claims data, specifically for dental providers. Medicaid is another U.S. program that provides health coverage to low-income people.[9] Their study involves multiple

[9] https://www.medicaid.gov.

analysis techniques and outlier detection methods based on specific metrics, such as number of unique beneficiaries and claim payments. They employ three univariate methods which include linear regression, box plots, and time series plots, as well as one multivarate method via clustering. They provide one case study using expected reimbursement values versus actual reimbursements with deviations, or outliers, indicated at 2.33 standard deviations away from the underlying regression model. The study incorporates 500 dental providers and they 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, based on expert evaluation. Another study, by Hu et al. (2012), 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 (Regression Trees, Random Forest, and MARS), then use Grubb's test to find outliers in the expected versus actual values. The purpose of Grubb's test is to assess the acceptable deviation ranges to find anomalous utilization levels. The authors demonstrate their method on 7667 diabetes patients, detecting 51 anomalies.

In Ekina et al. (2013) both healthcare fraud and Bayesian methods were incorporated. The authors use Bayesian co-clustering to identify potentially fraudulent individuals based on cluster memberships. Their focus is on healthcare fraud and the detection of unusual beneficiary-provider pairings. They target conspiracy fraud, which is fraud committed by more than one party. Their Bayesian model assumes Dirichlet priors for the marginal membership probabilities, and independent Beta priors for the Bernoulli random variable parameters. Samples are drawn from the posterior probability distributions to infer co-clusters of providers and beneficiaries based on latent variable probabilities. These posterior distributions are used to flag potential fraud activities via unusual cluster memberships. The authors test their algorithm on simulated data.

Even though research on both outlier detection and fraud detection are fairly ubiquitous, outliers are still very relevant issues today. Our study seeks to provide a better technique for outlier detection. We differ from the related works in several significant ways. Two of the studies (Trnka 2010; Ekina et al. 2013) use simulated data in order to test their methods, thus have limited value when comparing with real-world data and results. The study in Hu et al. (2012), while using patient utilization data, employs only Grubb's test for outlier detection. 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. Additionally, this test iterates over each value where multiple iterations can change the probabilities of detection. Furthermore, there is limited use of multivariate analysis or modeling in capturing potential fraud across multiple, possibly significant, variables. In our study, we use real-world Medicare data to demonstrate the efficacy of our multivariate Bayesian outlier detection model, which incorporates a probabilistic programming approach. The work presented herein is further differentiated from related works in its development and use of a multivariate regression model in conjunction with a probability model to create a generalizable outlier detection method for healthcare fraud, waste, and abuse.

[10] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm.

3 Current outlier detection methods

We briefly discuss several commonly used outlier detection methods for which we provide comparative analyses in Sect. 7. This paper does not provide a comprehensive discussion on outlier detection methods, so we refer the reader to Aggarwal (2015) and Zhang (2013). An outlier is simply some value that lies outside of the main grouping of data, or data points that do not naturally fit within some normal model (where a model can be based on distance, density, regression, etc.) (Aggarwal 2015). A trivial way to detect and flag outliers is to simply provide an upper and/or lower threshold value. For instance, given a vector of test scores between 0 and 100, one could apply a threshold of 90 for a great score and a 50 for a poor score. This method can be effective but assumes knowledge of the entire dataset. In this section, we outline several more sophisticated univariate and multivariate outlier detection methods. The interested reader should refer to the respective citations for further information.

3.1 Univariate outlier detection

Univariate outlier detection is predicated on using a single variable only in order to find outlying values. For instance, an outlier could be any point more than *t* standard deviations from the mean. This can be effective, but one problem with this method is that both the mean and standard deviation are quite sensitive to the presence of outliers in the dataset. In this section, we discuss several more sophisticated univariate methods.

3.1.1 Local outlier factor

An effective way to detect outliers, and better handle variability, is to use density- or distance-based techniques. One such method is known as Local Outlier Factor (LOF) (Breunig et al. 2000). LOF is based on local density, where locality is given by the *k*-nearest neighbors, whose distance is used to estimate the density. The local density is based on the distance at which a point can be reached from its neighbors, i.e. the reachability distance from a single point to its neighbors. An average ratio of the points reachability and its *k*-nearest neighbors' local densities is used as the LOF score. A disadvantage of this method is that these resulting scores are quotient-values and hard to interpret. A score of one or less indicates a normal value, but there is no clarity as to when a point should be considered an outlier. For instance, a score of 1.1 could be an outlier, but is hard to discern from only the LOF score. With that said, in practice, the scores can be significantly larger than 1, as well as having numerous scores above 1, making it difficult to come up with an outlier threshold[11]Breunig et al. (2000).

3.1.2 Grubb's test

The Grubbs' test is based on Grubbs' procedures for detecting outliers (Grubbs 1950, 1969). This test is defined by the null hypothesis that there are no outliers and the alternate hypothesis that at least one outlier exists in the dataset. The test detects one outlier at a time, though multiple iterations can change the probabilities of detection. Additionally, the use of Grubbs' test to find acceptable utilization deviations assumes

[11] https://turi.com/learn/userguide/anomaly_detection/local_outlier_factor.html.

normality, which may not be applicable to some datasets requiring data transformation or using a different test for outliers.[12]

3.1.3 Chi-Squared test

The Chi-Squared test (Tallarida and Murray 1987) is a hypothesis test where the sampling distribution of the test statistic is a Chi-Squared distribution given a null hypothesis test that is assumed true. This is done by creating a histogram of the observed data, and by compare these observed frequencies with a theoretical, or expected, frequencies. More specifically, Chi-Squared scores, which indicate outliers, are the squares of differences between values and mean divided by the variance. For any values between these quartiles, scores are always equal to zero, indicating an exact fit, otherwise they could indicate outliers.

3.1.4 Median absolute deviation

Instead of relying on the mean and standard deviation, the Hampel indicator (Ben-Gal 2005) replaces the mean with median and the standard deviation with the Median Absolute Deviation (MAD) scale. This is another threshold-based technique where the MAD metric is defined by Eq. 1. The outlier threshold is given by the differences between each value and the median, divided by the median absolute deviation. MAD is generally more effective than the other threshold-based methods, because it is less prone to the effects of outliers in the dataset.

$$\begin{aligned} MAD &= median(|x_i - median(x)|) \\ Threshold &= (x_i - median(x))/MAD \end{aligned} \tag{1}$$

3.1.5 Interquartile range

The standard boxplot rule (Williamson et al. 1989) uses the quartiles of a dataset to create specified thresholds and is less subject to the effects of outliers on the dataset. More specifically, the upper and lower bounds, which indicate possible outliers, are calculated via the interquartile range, as seen in Eq. 2. The value c is typically 1.5 or 3.0 to indicate outliers.

$$\begin{aligned} Threshold = (Q_1 - c \times (Q_3 - Q_1), \\ Q_3 + c \times (Q_3 - Q_1)) \end{aligned} \tag{2}$$

3.2 Multivariate outlier detection

In many cases, such as detecting fraud, using more than one variable can provide valuable information in detecting and investigating possible outliers. When compared to univariate methods, the number of techniques using multivariate input data to detect outliers is small. One reason to look at multivariate outliers, and regression models, versus univariate is that univariate outliers may not be extreme in the context of multiple regression, and a

[12] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm.

multivariate outlier may not be detectable in a two-variable or one-variable analysis. In this section, we briefly discuss two common multivariate techniques used to discover potential outlying data points.

3.2.1 Mahalanobis distance

This technique considers the scale of the data from many distributions (i.e. multivariate) expressing the probability of an observation. The Mahalanobis distance is similar to the Euclidean distance except that it standardizes the data along uncorrelated directions (Aggarwal 2015). This distance accounts for the variance at each variable and the covariance between variables, which can be considered the distance each observation is from the center of all the variables' distributions, or the centroids in a multivariate space. The Mahalanobis method gives the distance from a case to the centroid of all observations for the predictor variables (independent variable). A large distance indicates an observation that is an outlier in the space defined by the predictors (Stevens 1984). The general idea is to transform the data into standardized uncorrelated data and compute the ordinary Euclidean distance from the transformed data, from which the distances are calculated. Equation 3 defines the Mahalanobis distance.

$$M = (Y_i - \bar{Y})^{\mathsf{T}} \times S^{-1} \times (Y_i - \bar{Y}) \tag{3}$$

where Y_i is the design matrix of m variables and n instances (or observations), $\bar{Y}$ is the mean across the instances (a vector of size m), and S is the variance-covariance matrix of size $m \times n$. Although the Mahalanobis distance method can work with any dataset, it behaves best with data that are approximately multivariate normal. If the data are not multivariate normal, the averages might not be good representations of the center of the data and/or the general trends in the data will not be identified accurately using variance as a measure of spread. This can lead to inaccurate designations of outlying values.

3.2.2 K-means clustering

Partitioning of a data space results in k distinct clusters. One popular method of performing this is k-means clustering (Witten and Frank 2005; Robinson 2015). This method takes n observations, across m variables, and partitions them into k clusters, where each observation belongs to the cluster with the nearest average center, or centroid. The goal is to assign a cluster to each data point. K-means aims to minimize the distance from the data points to the cluster, or reduce the intra-cluster variance. It does this by minimizing the squared error function, as seen in Eq. 4.

$$J = argmin \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^j - c_j\|^2 \tag{4}$$

where J is the objective function, k is the number of clusters, n is the number of observations, x_i is the observation i, and c_j is the centroid for cluster j. Because k-means tries to minimize the within-cluster sum of squares, it always gives more weight to larger clusters. Additionally, k-means makes some assumptions on the data that include the following:

- The variance of the distribution of each variable is spherical (implying k-means wants spherical groupings of points around the mean when minimizing the within-cluster sum of squares)

- All variables have the same variance, or the spread of the clusters is similar
- Each cluster has roughly equal number of observations

If any of these assumptions do not hold, k-means will not perform as expected producing misleading or incorrect results. This can clearly be detrimental when trying to assess outliers for possible illegitimate activities. We chose k-means clustering for our study due to its popularity, ease of implementation, and usefulness on various types of data. Future work should consider other clustering models, such as hierarchical or spectral.

3.3 Normality assumptions

As mentioned, another important factor when looking at various outlier detection methods, is the shape of the data distribution. Generally, most of the common outlier detection techniques discussed will perform well on normally distributed data. Two of the described methods are noted as being dependent on normal distributions. Within the univariate methods, the Grubb's test, as noted in http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm, assumes normality thus requiring the data to either be normal or transformed, via a log or square root transformation for example. Out of the two multivariate techniques discussed, the Mahalanobis distance assumes a multivariate normal distribution (Cousineau and Chartier 2015; Aggarwal 2015). Again, applying a transformation function to the data can help create a multivariate normal distribution, but transformations do not always work and any results must be converted back to the original space. This weakness in these methods can lead to incorrect outlier detection, especially in non-normally distributed datasets such as the Medicare claims data.

4 Multivariate outlier detection method

In this section, we detail the two necessary components of our proposed outlier detection approach, which are the MARS regression model and our probability model, also known as a probabilistic program. Section 4.1 describes the MARS model and studentized residuals, which are used by the probability model to detect outlying values. For our probability model, we provide requisite background information on probabilistic programming and Bayesian inference in Sect. 4.2, and then discuss our probability model implementation in Sect. 4.3.

4.1 MARS regression model

MARS is a non-parametric regression model that accounts for the non-linearities between variables and their interactions (Friedman 1991). 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.[13] Equation 5 depicts the model built using MARS.

[13] https://documents.software.dell.com/statistics/textbook/multivariate-adaptive-regression-splines.

$$\hat{f}(x) = \sum_{i=1}^{n} c_i B_i(x) \tag{5}$$

where c_i is a constant coefficient, with a basis function $B_i(x)$ that is either 1, given just one term in the model, or $max(0, x - c_i)$ or $max(0, c_i - x)$ which is the hinge function. The advantages of MARS, which are beneficial in creating multivariate models, include the ability of the hinge function to automatically partition the input data (which somewhat contains the effects of outliers in the input data) and fast predictions. Furthermore, the automatic feature selection chooses the most relevant feature(s), thus reducing noise and possible masking when using the model outputs to detect outliers with many variables. Comparisons with other multivariate and/or nonparametric models is an option for future work.

The output from the MARS model is the residuals, or model errors, which are used to detect possibly anomalous activities. The absolute residuals from the model are the difference of the actual and predicted values, or $\varepsilon_i = y_i - \hat{y}_i$. In our study, we calculate internally studentized residuals (Stevens 1984; Mínguez et al. 2012) which is a way of normalizing the output of the residuals in order to directly compare residuals. Typically, the true standard deviation of the residuals is not known, so the estimated standard deviation is used to calculate the studentized residuals. These studentized residuals are a form of a Student's t-statistic, with the error estimate varying between points. Equations 6, 7, and 8 describe the process to generate internal studentized residuals, assuming an input design matrix X.

$$H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}} \tag{6}$$

$$\sigma^2 = \frac{1}{n - m} \sum_{j=1}^{n} \varepsilon_j^2 \tag{7}$$

$$t_i = \frac{\varepsilon_i}{\sigma\sqrt{1 - h_{ii}}} \tag{8}$$

where H is the hat matrix which is the orthogonal projection onto the column space of the design matrix X, ε is the absolute residual, σ^2 is the variance of the residuals, h_{ii} are the values in the diagonal of the hat matrix (also known as influence), and t is the studentized residual. For our method, we leverage these multivariate MARS model residual results to discover possibly fraudulent activities, for a particular specialty.

4.2 Probabilistic programming and Bayesian inference

For our research, we incorporate probabilistic programming (Gelman et al. 2014; Brooks et al. 2011; Davidson-Pilon 2015; Carpenter et al. 2016) to detect outliers in Medicare claims data. Probabilistic programming utilizes a high-level language to create probability models and solve them (via statistical inference) automatically. It is increasing in popularity and importance, in large part, due to the 2013 probabilistic programming initiative through the Defense Advanced Research Projects Agency (DARPA).[14]

In our study, we leverage probabilistic programming to perform full Bayesian inference. We provide a summary of some of the main points of Bayesian inference, with more

[14] http://www.darpa.mil/program/probabilistic-programming-for-advancing-machine-Learning.

detailed information available in Box and Tiao (2011). The main difference between Bayesian and more traditional statistical inference, the so-called frequentist view, is the preservation of uncertainty. The view is that probability, in a Bayesian sense, is a measure of belief, or confidence, of an event occurring. Bayesian inference provides a way of combining new evidence with prior beliefs, or assumptions, through the application of Bayes' rule, defined in Eq. 9.

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)} \tag{9}$$

In Bayes' rule, P(A) is the prior belief in event A (previous assumptions or beliefs based on some prior knowledge), P(X) is the prior probability of the evidence, P(X|A) is the likelihood of evidence X given event A (a single value X for a hypothesis A), and P(A|X) is the posterior probability, which is the updated belief based on the evidence. This equation essentially takes our prior knowledge about the parameters and updates this knowledge with the likelihood to observe the data for particular values generating the posterior probability. The posterior is the probability of a value given the data and our prior knowledge. Additional information on Bayesian inference and probabilistic programming can be found in Appendix 1.

4.3 Probability model

To create our outlier detection probability model, to detect outlying values from the MARS regression model residuals as part of our twofold approach, we use the probabilistic programming language known as Stan (Carpenter et al. 2016). The posterior distributions, via Stan, are drawn from the full conditional of each unknown parameter. This is done using Hamiltonian Monte Carlo (HMC) and the No-U-Turn Sampler (NUTS), which are both implemented in Stan to perform statistical inference. The model fitting is done by specifying the full likelihood function and the prior distributions of all unknown parameters. This study is not a tutorial on Stan, but the interested reader can find additional information in Carpenter et al. (2016), Gelman et al. (2014), Savage (2016) and Stan Development Team, Stan Modeling Language Users Guide and Reference Manual. In the remainder of this section, we provide details on the proposed probability model.

We declare three input variables that correspond to the actual inputs (see Appendix 1 for the model code) including the vector for the population (or sample) from which to assess outliers, the values to check in order to find outliers, and the length of each vector. The unknown model parameters for Stan to estimate are the mean and standard deviation of the sample space from the inputted values to check (from which to detect outliers), and degrees of freedom for the Student's *t* distribution (the likelihood distribution). For the detection of outliers, we assume a Student's *t* distribution for the probabilities, using mean, standard deviation, and degrees of freedom as defined in Eq. 10, showing the probability density function.

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}} \tag{10}$$

The *t* distribution (Ross 2002) is symmetric and bell-shaped, similar to the normal distribution with a peak at zero. The difference is that the spread of the data is more than that of the standard normal distribution. A *t* distribution would thus better capture wider tails in a dataset in order to better describe the outlier distribution. This is like the approach taken in Chaloner and Brant (1988), with additional information found in Savage (2016).

We define non-informative, or vague, priors of normal distributions for both the mean and standard deviation, and a Cauchy distribution for the degrees of freedom (Gelman et al. 2014; Savage 2016). This is because the sample standard deviation is used and is a random quantity varying with various samples, thus larger variability and spread in the t distribution. From these, the likelihood of the value is defined as a t distribution constrained at a lower bound of zero. As mentioned with the Bayesian approach, evidence is likely to either support or supplant any prior assumptions; thus, the use of non-informative priors does not have a significant effect on the model likelihood as the majority of the probabilities are more influenced by the evidence rather than the prior assumptions. Finally, we declare the quantities other than the simulated parameters to be generated. In this case, we simply generate cumulative t distribution functions for being both greater than and less than some value, where the t distribution cumulative distribution function (cdf) is $\int t(v)dx$, with variables from the generated quantities Stan model block (see Appendix 1).

In using the Stan probabilistic programming language, probability statements are written in a standard notation, e.g. $y \sim Normal(\mu, \sigma)$ (Savage 2016, Stan Development Team, Stan Modeling Language Users Guide and Reference Manual). This means that the variable y modeled data is declared as a normal distribution with some given mean and standard deviation. We use this notation to express our probability model. Equation 11 defines the three prior distributions for the parameters (one for each of the following: μ, σ, and v), and Eq. 12 shows the data model (likelihood distribution) given those parameters. Note the Cauchy distribution is restricted to only positive values. The parameters within each *Normal* and *Cauchy* prior distribution function (μ_1, μ_2, σ_1, σ_2, x, and γ) are fixed values that summarize any prior information that were chosen based on data exploration and experimentation.

$$\begin{aligned}\mu &\sim Normal(\mu_1, \sigma_1)\\ \sigma &\sim Normal(\mu_2, \sigma_2)\\ v &\sim Cauchy(x, \gamma)\end{aligned} \tag{11}$$

$$y \sim \text{Student's } t(v, \mu, \sigma) \tag{12}$$

For the final result, we compute the probability of observing a more extreme value versus the majority of observations, which is stored in the outlier probability model variable *prob*. To interpret this probability, we observe that a probability of 50% says that there would be a 50% chance of seeing a value more extreme than the current value. This would be akin to a value in the middle of a typical normal distribution or t distribution "bell curve", thus unlikely to be an outlier. In contrast, a probability of 1% would indicate only a 1% chance of seeing a more extreme value, which is a probable outlier as there are not many values more extreme. This is similar to being at the tails of the distribution. These thresholds for the probability of being an outlier are data dependent and should be considered carefully when identifying and investigating outlying values. With that said, some possible ramifications of having suboptimal thresholds include increased numbers of false positives or false negatives.

5 Medicare data

The data that the Centers for Medicare and Medicaid Services[15] has released, at the point of this publication, are for calendar years 2012, 2013, and 2014. We use the Physician and Other Supplier Data 2012–2014 dataset, which describes payment and utilization claims data for Medicare Part B services with information on services and procedures provided to Medicare beneficiaries. Due to the large size, we decided to limit the data to office clinics in Florida (as opposed to larger facilities, such as hospitals). Because of Florida's unique demographic, this subset is not necessarily representative of the entire US population. However, Florida is a good candidate for our study in having the second highest number of Medicare beneficiaries and being second in total Medicare spending.[16] Table 1 summarizes both the complete and Florida-only Medicare datasets, where the columns indicate the following: the number of instances is the count of all combinations of procedures performed by providers, the number of features are the variables, the unique providers column is the count of physicians, and the distinct procedure types are the various unique procedures performed.

The Medicare dataset contains values that are recorded after claims payments were made and with that, we assume that the Medicare dataset was appropriately recorded and cleansed by CMS,[17] thus are not concerned with bad values adversely affecting the regression model. We do not remove values prior to model creation, since these input values are all considered valid points for possible fraud detection. The 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)[18] code, and National Provider Identifier (NPI).[19] 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) to create and file a claim. Based on this information, including the HCPCS codes, Medicare processes the claim and makes the appropriate payment.

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,[20] thus are actual provider services versus drug-specific activities and/or prescriptions. From the 26 total variables, we selected 13. Out of these 13 variables, 7 are used for the regression models, as seen in Table 3, while the other variables, shown in Table 2, can be used for the identification of possibly fraudulent activities and corresponding provider(s). The

[15] https://www.cms.gov/research-statistics-data-and-systems/research-statistics-data-and-systems.html.

[16] http://kff.org/state-category/medicare/.

[17] https://questions.cms.gov/.

[18] https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/.

[19] https://nppes.cms.hhs.gov/NPPES/.

[20] https://questions.cms.gov/.

Table 1 2012–2014 Medicare physician and other supplier data summary

Dataset	Number of instances	Number of features	Unique providers	Distinct procedure types
United States	27,757,455	26	1,049,362	6741
Florida	1,197,238	26	48,230	2922

Table 2 Description of medicare variables for identification

Variable name	Description
Provider Type	Medical provider's specialty, e.g. Cardiology
HCPCS Code	Code for specific medical services furnished by the provider
NPI	Unique provider identification number
First Name	Provider's first name
Last Name	Provider's last name
City	Provider's office city

remaining 13 unused variables are not currently incorporated into our method, and include features such as the free text, written by a provider, describing an office visit. The use of these remaining variables, along with applying different feature engineering approaches, is left as future work.

It is important to note that all variables used for the regression model are uniquely defined by Medicare,[21] thus are considered fair and valid predictors. To further validate this uniqueness, we compared *Avg. Medicare Payment Amount* and *Avg. Medicare Allowed Amount*,[22] per specialty, via two hypothesis tests to statistically characterize the differences between these distributions. Because of the skewed nature of the Medicare amount distributions, we used unpaired Mann–Whitney and Kolmogorov–Smirnov tests.[23] Both tests indicated a significant difference between the two distribution at a 95% confidence interval. The remaining variables also have similarly different distributions.

Additionally, the *Zip Code* and *Year* variables are encoded as factors, which provide for categories for all of the values in each variable.[24] This is akin to creating groupings or categories for each of the zip codes and years. We encode these two variables as factors to better represent localization of provider utilization and services, since the Medicare data does not explicitly account for these variations. Thus, these two variables are not considered numeric values when passed into the regression model. The remaining variables are kept as numeric values.

Furthermore, we limit the experiments in this initial study to 7 of the 83 provider types (specialties), which include Thoracic Surgery, Cardiology, Ophthalmology, Optometry,

[21] https://questions.cms.gov/, https://www.healthcare.gov/glossary/.

[22] https://secure.wpsic.com/sales-materials/files/28807-medicare-approved-amounts-tip-sheet.pdf, http://www.nolo.com/legal-encyclopedia/part-b-medical-insurance-what-you-pay.html.

[23] www.minitab.com.

[24] https://www.R-project.org/.

Table 3 Description of medicare variables for regression model

Variable	Mean	Median	SD	Min	Max	Description
Zip Code	–	–	–	–	–	5-digit code
Year	2013	2013	–	2012	2014	2012, 2013, and 2014 Medicare years
Line Service Count	218.90	60	599.23	11	40,569	Number of services provided/ procedures performed per provider
Total Count	1487	1168	1231.16	1	4341	Sum of procedures performed across providers
Beneficiary Day Service Count	185.60	57	386.04	11	18,037	Number of distinct Medicare beneficiary per day services
Avg. Medicare Allowed Amount	$131.95	$82.46	$384.84	$0.01	$17,022.25	Allowed amount for the services (with deductible and coinsurance)
Avg. Medicare Payment Amount	$99.91	$59.22	$301.96	$59.22	$13,341.32	Amount Medicare paid the provider for services performed

Table 4 Provider type dataset summary

Provider type	Number of instances	Number of providers	Number of distinct services	Average payment amount
Thoracic Surgery	1136	129	64	$110.79
Cardiology	76,105	1685	559	$125.77
Ophthalmology	44,552	1226	250	$84.83
Optometry	22,989	1568	79	$60.05
Dermatology	57,399	947	295	$130.15
Psychiatry	8123	938	172	$67.61
Nurse Practitioner	35,838	3699	656	$47.88

Dermatology, Psychiatry and Nurse Practitioner. In choosing the provider types, or specialties, for our study, we selected them 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 4 summarizes the provider type datasets.

6 Experimental design

Our method incorporates two parts in order to detect outliers and possible anomalous behaviors. The first part involves a nonparametric multivariate regression model to capture the the most amount of information found in all the possible predictors. We create models based on the publicly available Medicare dataset, and specific medical specialties, as described in Sect. 5. Because we are interested in detecting anomalous behaviors in this Medicare claims data, specifically using average Medicare payments as the response variable, we split our Florida-only dataset into distinct groups by provider types (or specialties) and HCPCS procedure codes. This grouping allows us to easily produce distinct modeling of payment data by medical specialties, as described in Table 4.

We decided on the MARS model based on its high performance, robustness, and its ability to provide nonparametric, multivariate model fit. We use the R programming language,[25] with the earth package (Milborrow et al. 2016) implementation of MARS, to create and validate each model. The Classification And REgression Training (CARET) (Kuhn et al. 2016) package is used to create the final MARS model, with 10-fold cross-validation. We chose 10-fold cross-validation to reduce overfitting of trained models. Our use of 10-fold cross-validation incorporates 90% of the dataset for model training and 10% for testing, within each fold. This provides us a trained model using most of the data, thus able to closely estimate the prediction performance of the final model using 100% of the available data. 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. The MARS model, and corresponding residuals, is the first part in detecting Medicare claims payment outliers addressing the multivariate nature of the dataset. Table 5 lists the 10-fold cross-validation model training results, per medical specialty. The metrics shown include the following:

- Mean Absolute Error (MAE)—Average of the absolute differences of predicted and observed values (lower is better)
- Root Mean Squared Error (RMSE)—Estimated standard deviation of unexplainable variations in the dependent variable (lower is better)

Given the 7 predictors used to create each model, these results indicate a good model fit for each specialty. The better model fits for specialties, such as Dermatology, Cardiology, and Thoracic Surgery, appear to indicate that these specialties are less general in the types of procedures performed (not to be confused with the number of procedures performed), thus have a more specific and homogeneous nature in the procedures performed. Additionally, the absolute range of Medicare claims payment amounts vary greatly between and within each specialty, lending to further model performance fluctuations.

Once the multivariate regression models are created for each specialty, the second part of our method involves the detection of outlying values using the resultant studentized model residuals. To create our fully Bayesian outlier detection probability model, we used the rstan (Guo et al. 2016; Carpenter et al. 2016) implementation of the Stan probabilistic programming language. Each Stan probability model was run with 4000 iterations and 2 Markov chains (Brooks et al. 2011) (for performance reasons). In order to assure convergence, we also adjusted the *adapt_delta* parameter to 0.999 (per the Stan message after running the model with the default value of 0.8), which is the target acceptance probability. Correspondingly, the step size and tree depth needed to be adjusted as well to 0.001 and 20, respectively. This did slow the modeling process, but increased convergence giving a split-$\hat{R}$ multi-chain diagnostic for all parameters of less than 1.1 for each value, over all Markov chains (Gelman et al. 2014). If the split-$\hat{R}$ is larger than this, then the model results are not reliable and should not be used. Furthermore, as is typical with outlier detection methods, the input data for the Stan probability model has identical vectors for the *value* and

[25] https://www.R-project.org/.

Table 5 MARS model training metrics by specialty

Specialty	MAE	RMSE
Ophthalmology	3.855	6.423
Thoracic Surgery	3.568	5.810
Optometry	4.183	6.248
Dermatology	3.549	5.399
Psychiatry	5.333	7.677
Cardiology	3.626	11.65
Nurse Practitioner	4.066	6.462

check_value variables (from the Stan model code in Sect. 4.3). This implies that we are looking for any and all outliers in the full population (3 years of Medicare data in our study).

The following sections outline the analyses to assess and compare our outlier detection model against both other univariate and multivariate methods. In each section, we briefly discuss the configuration for the other outlier detection methods. We intend to show that our method provides comparable results and more information on the probability of an outlier versus common univariate methods run against our multivariate regression residuals, as well as our method's performance versus two common multivariate outlier detection methods.

7 Comparisons with existing methods

7.1 Univariate outlier detection comparison

For this comparison, we take the MARS model studentized residuals and compare our fully Bayesian probability model against five other univariate outlier detection methods. We aim to show that our method provides more consistent results and probability information that is not found with the other methods. In order to be consistent and provide clearer context for outliers, and possible anomalous activities, we decided to use the top 1% of outliers found by each method. Note for our method, these are actually the outlying values with a probability of 1% versus simply the top 1% of the ordered outliers from the other outlier detection methods. Our approach in selecting the top *n*% of values with the highest outlier scores as anomalies has been employed previously using LOF to flag outliers.[26] Even so, there are various methods that exist, in practice, for selecting outliers based on scores with no single method used exclusively in application. We leave the comparison of different outlier detection thresholds as future work.

For LOF, several values of *k* were tried in order to optimize the discovered outliers. We decided to use 200 neighbors, which was able to capture an appropriate number of outliers across all tested specialties. Even so, the number chosen is not necessarily optimal, as there is no effective or efficient way to dynamically choose the correct number of neighbors for a specific dataset or type of data. Given how LOF functions, this is a limitation in detecting outliers, implying a higher likelihood of too many false alarms or missing too many real anomalous activities. A two-sided Grubbs' test was also run, but the algorithm is, by

[26] https://turi.com/learn/userguide/anomaly_detection/local_outlier_factor.html.

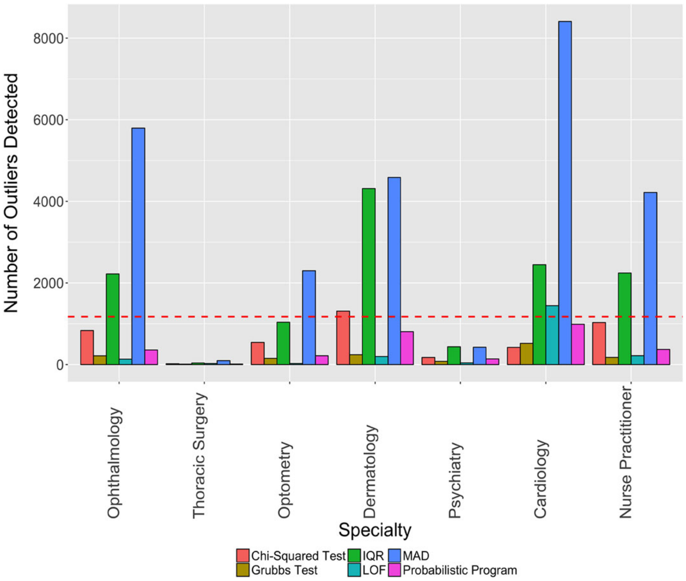

Fig. 1 Detected outliers by specialty and method (Color figure online)

default, set to find either an outlier on each tail (high or low outlier only) or the most extreme outliers for both tails (both the highest and lowest outliers). In order to bypass this limitation, we implemented a “hold-out” iterative process. First, we find the most extreme outlier (either high or low), remove this outlier from the dataset, then continue the iteration on the updated dataset minus that found outlier. This then allowed us to find outliers across the entire dataset based on the Grubb’s test algorithm, from which to take to top 1% to compare with the other outlier detection methods. Finally, the Chi-Squared Test, Median Absolute Deviation, and Interquartile Range outlier detection methods leverage the outliers R package (Komsta 2015), which automatically determined outlier detection thresholds (see Sect. 3).

Figure 1 depicts the number of outliers found, by detection method, within each specialty (the average number is indicated by the red dashed line). Each method in our experiment returns a different number of possible outliers detected based on how the outlying values are flagged using a 1% probability threshold (since each method has a different way of determining outliers). For instance, the Chi-Squared test, described in Sect. 3, has an input of 1% for the probability, but finds the flagged outliers based on the Chi-Squared quantile function; and the MAD and IQR methods create a static threshold with anything above that line considered an outlier. The threshold-based methods, such as IQR and MAD, consistently find more outliers than the other methods. This most likely indicative of numerous false alerts, i.e. flagging of anomalous activities that aren’t really anomalous, that will require manual tweaking to the thresholds per specialty. Even with more optimal thresholds, the inclusion of new data or new specialties will necessitate threshold adjustments. The Chi-Squared method is fairly consistent, but has a high number of detected outliers and diverges (in direction) from all the other methods for Nurse Practitioner. The breadth and number of procedures in the Nurse Practitioner specialty is most likely the cause of this divergence. LOF and Grubb’s test are both consistent methods indicating lower numbers of outliers than all the other methods, including ours, but, as

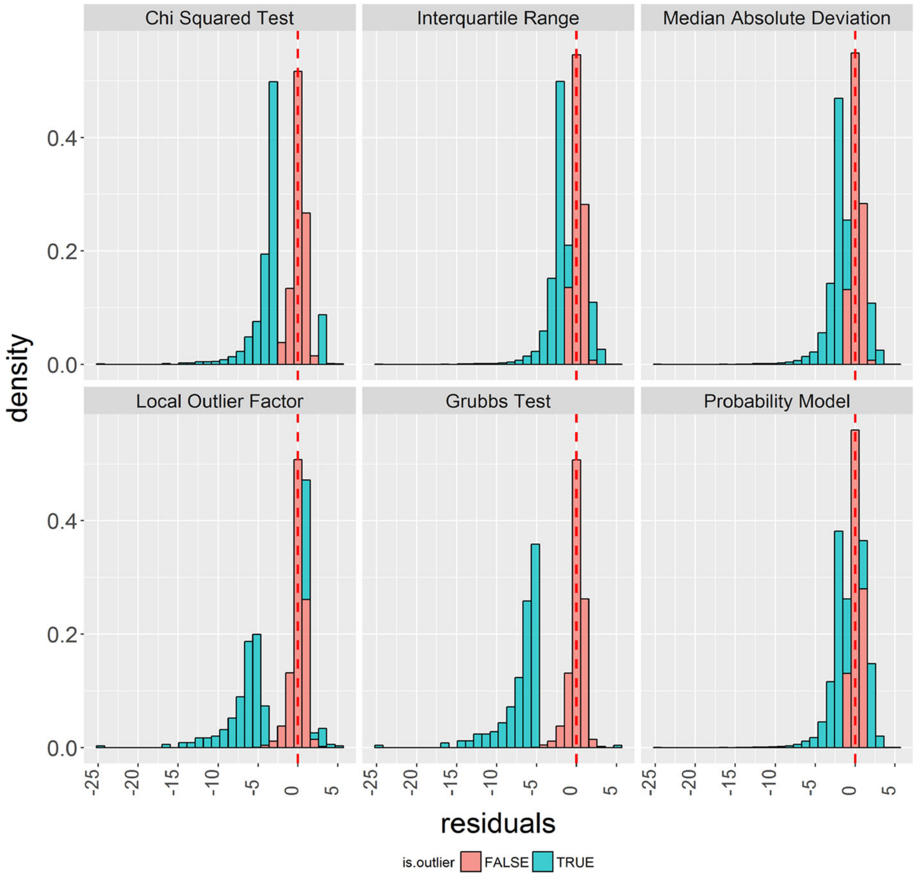

Fig. 2 Dermatology histograms by detection method

stated before, the major limiting factors are choosing the number of neighbors for LOF and the necessary construction of the "hold-out" iterative process and normality assumptions for Grubb's test. Therefore, even given these results, the limitations make it unclear as to whether these are the optimal settings and, thus, an appropriate number of detected outlying values.

When noting patterns in Fig. 1, there are some specialties where the methods give very similar results, but others that diverge substantially with the number of outliers detected. Dermatology, for example, has large differences between methods due to the number of procedures performed and the range of payment values across procedures. On the other hand, Thoracic Surgery has very few procedures and a tight range for procedure payments. This divergence in methods can be seen comparing the histograms in Figs. 2 and 3, by noticing the shift in densities for both outlying and non-outlying values. Even though Thoracic Surgery is more sporadic (due to having less observations), it is more consistent in density across the methods, especially with the detected outlier distributions.

As seen, our method is consistent in its detection of outliers in the Medicare dataset, being both less variable, in general, by specialty and not dependent on manual or static thresholds or neighbors. Moreover, our method provides information beyond what is

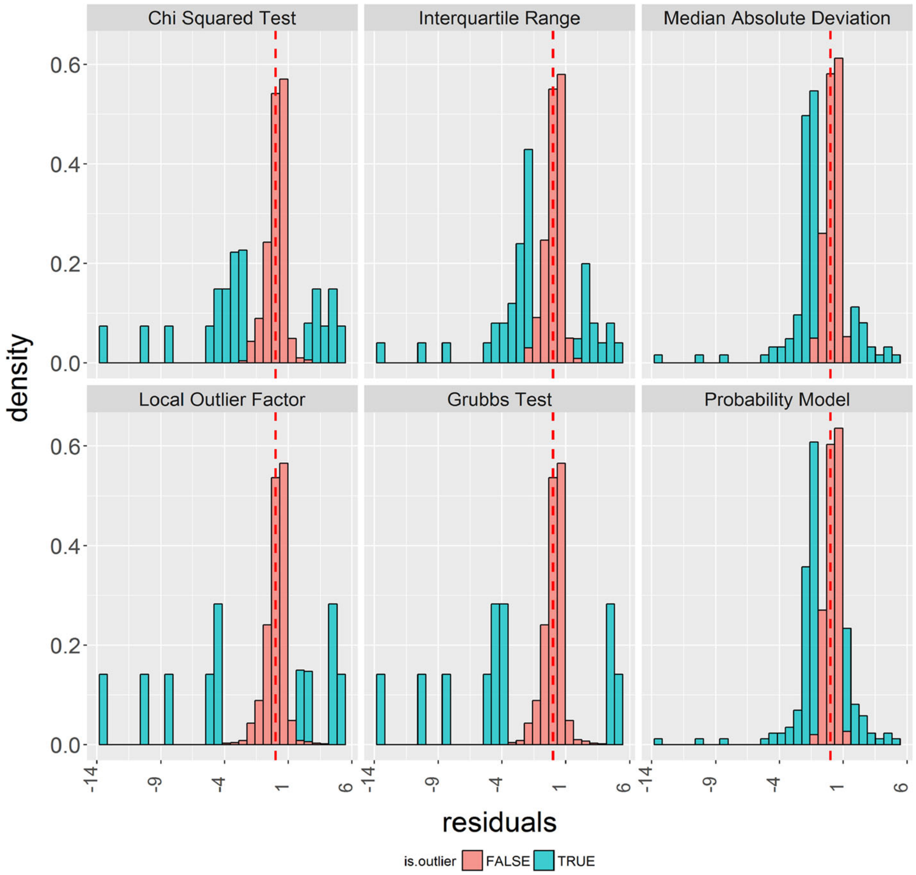

Fig. 3 Thoracic surgery histograms by detection method

available in the other outlier detection methods. These other methods provide point values or thresholds for possible outliers, but do not provide additional information such as probability of a value actually being an outlier. Even more sophisticated methods, like LOF or Grubb's test, do not provide this, but rather give scores that are subject to interpretation as to what constitutes an outlier.

Our model for outlier detection, which utilizes full Bayesian inference, provides a distribution of probabilities per value contributing information to help discern real outliers. An example of this extra information is found in the credible intervals, as seen in Fig. 4, which depicts a sample, directly from the Stan probability model, of 50 point value probabilities for Psychiatry. The y-axis shows the 50 probability model variables for each of the point values and the x-axis is the outlier probability. The black dots are the mean probabilities, with the red lines indicating the credible level probabilities and the black lines showing the outer level probabilities. These intervals provide a view on the distribution of probabilities per point value allowing for a more dynamic assessment of what is and is not an outlier. The thresholds to use, in considering anomalous activities, can be chosen by mean and/or credible interval depending on the specialty and the needs of the

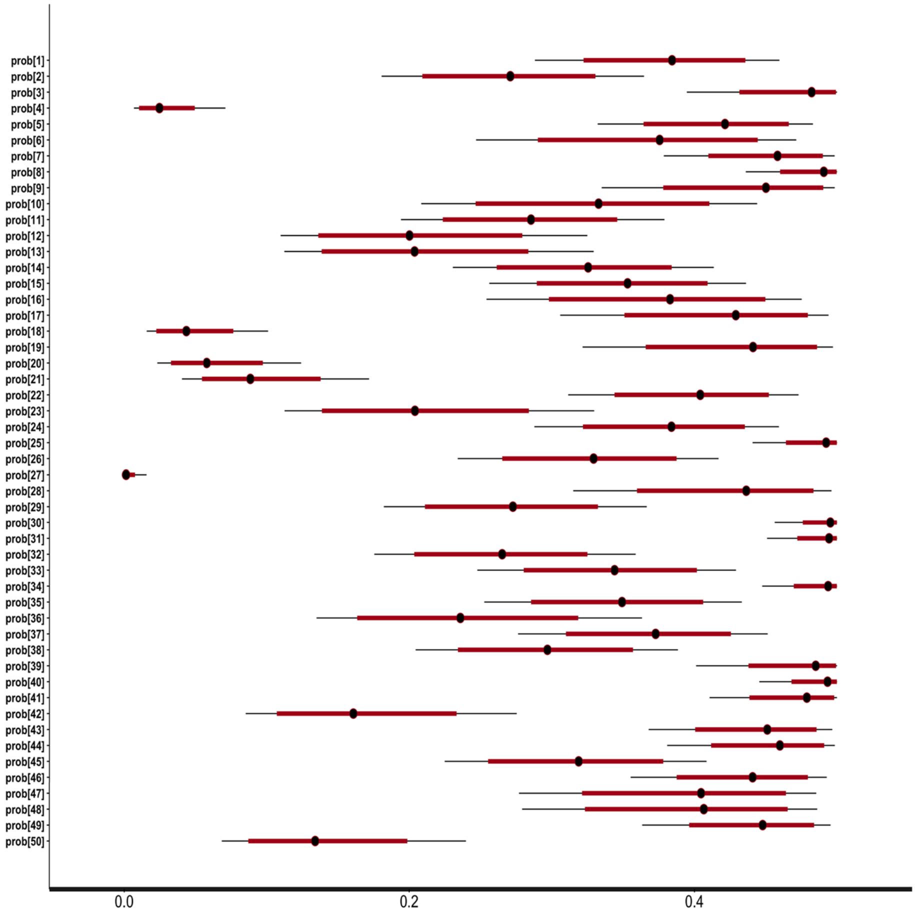

Fig. 4 Psychiatry point value outlier probabilities (Color figure online)

organization, e.g. choosing to look at 5% of outliers versus 1% based on expert knowledge, and considering intervals that overlap these thresholds as possible outlying values.

It is important to reiterate that this is outlier detection on real-world data, thus it is difficult to validate anomalous cases (unlike with artificial or toy data with predetermined outliers). Outlier detection is an unsupervised process that requires external inputs and further investigation to determine whether it is truly an anomalous activity or not. Thus, the detection of outliers is not a perfect process. Given this, our probability model provides more flexibility, over the other univariate outlier detection methods, in determining outlying values utilizing probabilities, and is not restricted by distribution assumptions or neighbor configurations. We discuss some validation in Sect. 8 in applying our method to the various specialties in the Medicare dataset.

7.2 Multivariate outlier detection comparison

Because our method incorporates two parts to provide multivariate outlier detection, it was necessary to compare various univariate outlier detection methods and demonstrate the validity and efficacy of our probability model. The aim of our method, though, is full multivariate outlier detection. In this section, we provide a comparison of Mahalanobis distance and k-means clustering versus our full method. We perform experiments with the Medicare data. For clarity, due to a smaller number of observations, we chose the Thoracic Surgery specialty and the top 1% of flagged outliers for each method.

When running Mahalanobis distance, we needed to determine the threshold, or distance, with which to consider values as outliers. For this comparison, we ordered the distances, in decreasing order, and manually chose the top 1% as outliers to compare. Clearly, this process is not extensible beyond a handful of specialties at a time. Moreover, the determination of a distance cut-off is somewhat arbitrary and highly dependent on the specialty chosen. The calculated distances do not intrinsically indicate meaningful information for each outlier, as would probabilities of being actual outliers. The detection of outliers using Mahalanobis can be done in several different ways. One way uses values such as 1.5 or 3.0 times the mean Mahalanobis distance (Davenport 2013) to create the threshold to compare against each of the distances (or scores), with a value above the threshold considered an outlier. Also, reference “http://www-01.ibm.com/support/docview.wss?uid=swg21480128” generate *p* values from a Chi-Square cumulative distribution function and use these pvalues to check against some threshold value, such as 1%. Finally, using a determined percentage of points that should be flagged as outliers, e.g. some probability of points that are to be considered outliers, from a vector of sorted scores has also been used as a threshold for detecting outliers (Rosenmai 2013). As mentioned, we follow the latter technique in flagging outlying values.

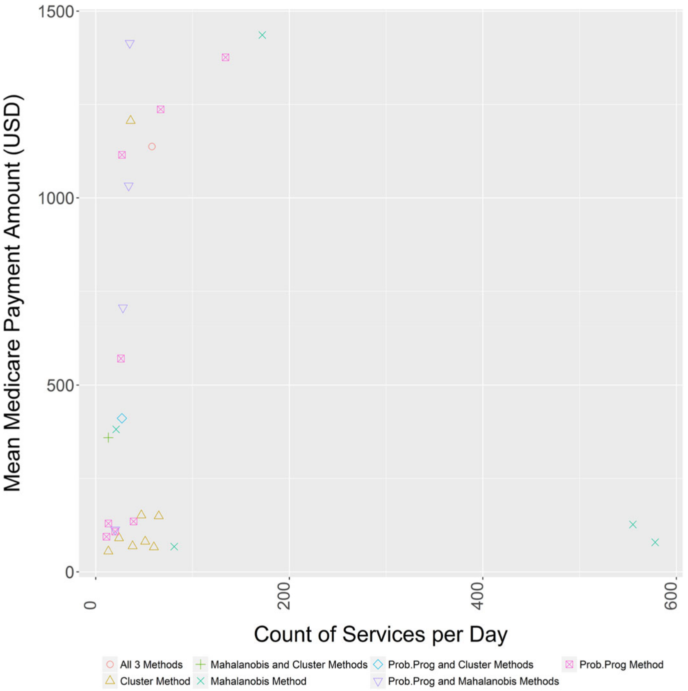

Fig. 5 Thoracic surgery multivariate outlier comparison

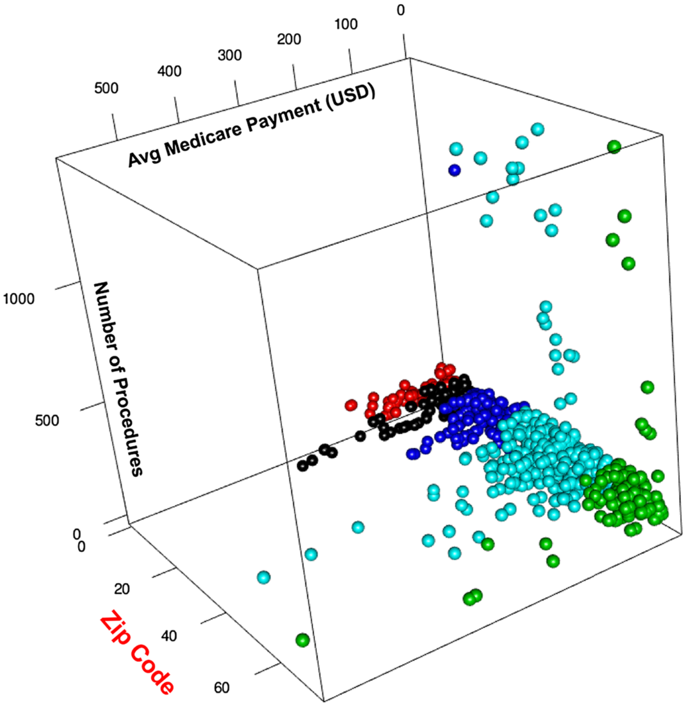

Fig. 6 K-means clustering zip code influence

For k-means clustering, the limiting factor is deciding on the number of clusters. For our analysis, we used the so-called "elbow method" which visually depicts the within-cluster variance versus the number of clusters. The optimal number of clusters is chosen when the variance reaches a stable point, i.e. at the elbow in the plot. The clustering process can be difficult due to the selection of the number of clusters and, even if a reasonable number of clusters is chosen, these clusters do not necessarily directly convey any meaning about outlying values, as opposed to probabilities and credible intervals.

Figure 5 shows the detection of outliers for all three methods (probability model, Mahalanobis distance, and clustering), by Medicare payment and count of services per day. For Thoracic Surgery, based on the elbow method, we chose 5 clusters. There is more overlap between our method and Mahalanobis distance, with clustering having the least overlap with the other methods. Additionally, clustering appears to focus on the low count, low payment area and not labeling any high payment values as outliers. The reason for this, is that the clustering appears to be heavily influenced by the zip code predictor. Figure 6 clearly depicts the clustering tendency towards zip code across the 5 clusters. The zip code clusters are the dominate groupings in both payments and procedures performed, with no noticeably distinct clusters for either of these other two variables.

Notice that the Mahalanobis distance method captured the high count per day, but no other method captured these values as outliers. It could be assumed that these are actual outliers because the number of procedures performed is higher than the rest, but these are more likely false alarms due to the multivariate normal requirement for Mahalanobis

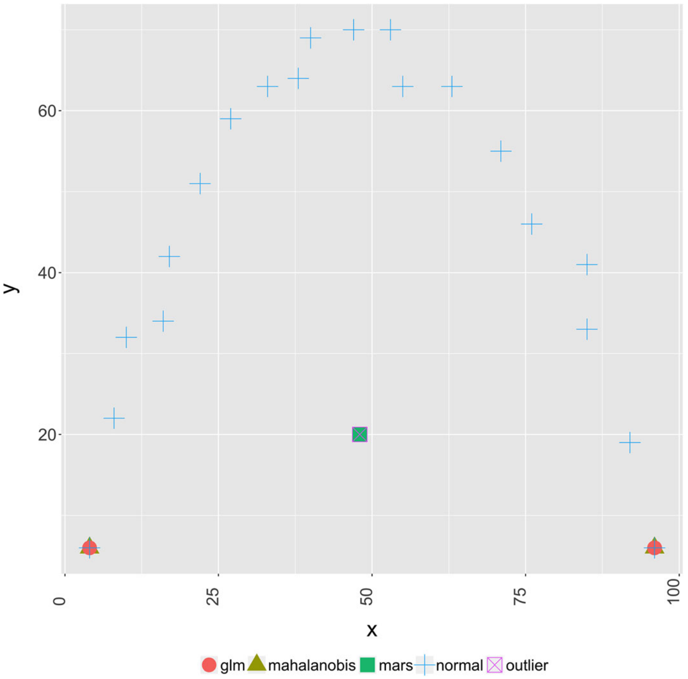

Fig. 7 Non-linear relationship outlier example

distance. Two tests were done in order to check for normality, the first was Anderson-Darling's Normality test[27] on all the individual predictors. The *p* values were all zero indicating each predictor's distribution as non-normal, at a 95% confidence. The second check involved Henze–Zirkler's Multivariate Normality and Mardia's Multivariate Normality tests (Korkmaz et al. 2014), both of which indicated the dataset was not multivariate normal. Furthermore, a toy two-variable example is shown in Fig. 7 where the *x* and *y* variables have a non-linear relationship. The outlier is obvious and clearly seen in the lower middle of the plot at location (50, 20). The MARS model, which captures nonlinear relationships, using our probability model, successfully finds the outlier, whereas running a generalized linear model (GLM) (Dobson and Barnett 2008), i.e. a multiple linear model, does not, showing the tails as outliers. Mahalanobis distance does not detect the obvious outlier, again flagging the tails as outliers. Moreover, the Mahalanobis distance method does not perform any better than GLM, both of which are unable to detect the single outlier.

For both the univariate and multivariate outlier detection comparisons, our method is comparable to the other outlier detection methods and, more critically, provides valuable probability distribution information per point value in assessing an outlier's validity and usefulness. In the following section, we employ our method to detect anomalous activities in Medicare claims payment data to flag possible fraudulent behaviors.

[27] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm.

8 Medicare data outlier detection analysis and discussion

As we have discussed and demonstrated, our method provides more meaningful results, when compared to other common univariate outlier detection methods. The values returned from our model are distributions of probabilities of a point value being an outlier. Thus, the mean probability of being an outlier can be used to determine the appropriate probability threshold needed to indicate an outlier, given some particular dataset. Unlike the other methods discussed in this paper, our method is based on the probability of an observation being an outlier, and is not dependent on a set threshold of distance values. Thus, if a value has a 5% probability, this implies that this value is significantly less similar to the other values; conversely, if the value is at 45%, then it is similar to the majority of values. The creation of these probabilities is done automatically, and inherent to the model, with no need to specify the number of neighbors or clusters, group or split the data, or create static and arbitrary cut-off limits to indicate an outlying value versus a normal one. Additionally, our method is non-parametric, therefore not limited to normal or multivariate normal distribution assumptions.

In addition to the average probability of being an outlier, our method produces credible intervals. As discussed, beyond simply providing a distribution of mean probabilities, our model generates credible intervals for each value showing that a particular value has an 80 or 95% probability of being within its probability distribution. Figure 8 shows the credible intervals for Thoracic Surgery, for probabilities below 5% only for visualization purposes. In these plots, the yellow dots indicate the mean probability, the red horizontal line is the credible level with 80% intervals, and the black horizontal line is the outer level with 95% intervals. This distribution of credible intervals helps determine how confident we are in our assumptions that a value is indeed an outlier. For instance, values with intervals that are clearly to the left or right of the 1% probability threshold can be confidently flagged as either an outlier or normal, respectively. If the interval crosses over the 1% threshold, this point could be considered an outlier rather than normal, or vice versa, depending on where the average probability lies in relation to the threshold. Figure 9 depicts the credible intervals for Cardiology. When comparing both Cardiology and Thoracic Surgery credible intervals, the intervals are wider for Thoracic Surgery indicating more variance in the dataset. This information captures inherent uncertainty and could be used to help create better indicators as to what constitutes an outlier given a more variable dataset or, conversely, assume tighter bounds when flagging outliers with a less variable dataset.

We apply our full methodology to the 7 Medicare specialties showing possible fraudulent activities, that could require further investigation, as well as two real-world, documented fraud cases for model validation. In general, the flagged fraudulent activity values can be associated with various characteristics such as NPI, last name, address, and zip code. These are readily available identification metrics that can be used to either ignore a possible fraud flag or investigate further. In this study, we use flagged value labels composed of a combination of the first name and masked NPI for privacy and confidentiality.

To support the effectiveness of our outlier detection method, based on the experimental results, we use the NPI numbers to search for known fraud cases in news reports and the US Department of Health and Human Service Office of Inspector General Exclusion List.[28] When searching for providers via last name and NPI, we discovered a Cardiologist under

[28] http://oig.hhs.gov/exclusions/index.asp.

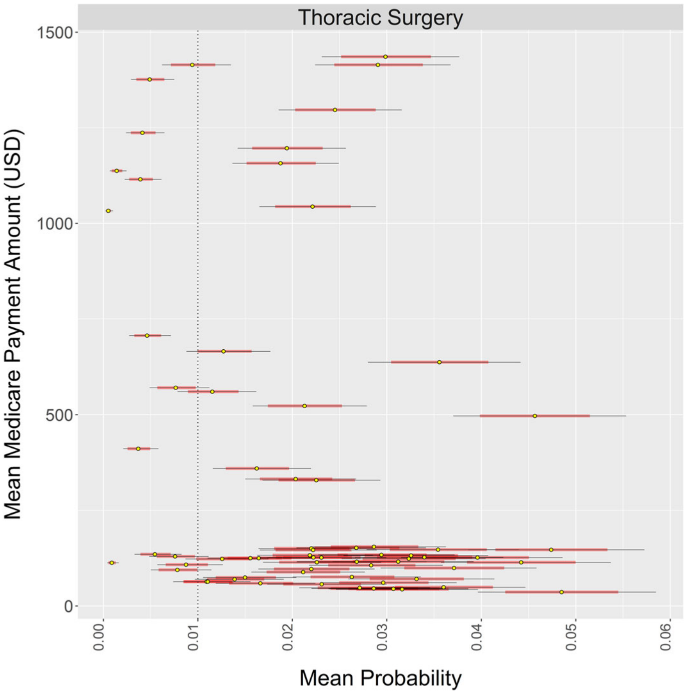

Fig. 8 Thoracic surgery credible intervals

investigation for fraud.[29] It was reported that this provider billed Medicare for medically unnecessary peripheral artery interventions, which is flagged by our method, with a label of *Asad:5487* in the town of Ocala, Florida, as seen in Fig. 10. The *Asad:5487* labeled probabilities are 0.02, 0.06, and 0.16 indicating these points are quite different from most of the other Medicare Cardiologists. 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. Additionally, we discovered an ophthalmologist under criminal investigation for alleged excessive billing of Medicare.[30] It was reported that this provider's Medicare payments went toward reimbursements for injections of a costly drug, Lucentis, to treat patients with macular degeneration, a retinal disease. Figure 11 shows this possibly fraudulent activity labeled with *Salomon:8371* from West Palm Beach, Florida, with probabilities 0.14 and 0.61.

Even given these successful detections, due to the limited number of real-world fraud cases pertaining to the current 2012–2014 Physician and Other Supplier PUF data, there is a gap in assessing performance with mostly anecdotal fraud investigation cases. Additionally, the publicly available Medicare claims data does not provide labels indicating

[29] http://www.ocala.com/article/20160422/ARTICLES/160429933, https://www.justice.gov/opa/pr/government-intervenes-lawsuit-against-florida-cardiologist-alleging-unnecessary-peripheral/.

[30] http://www.miamiherald.com/news/local/community/miami-dade/article1962581.html.

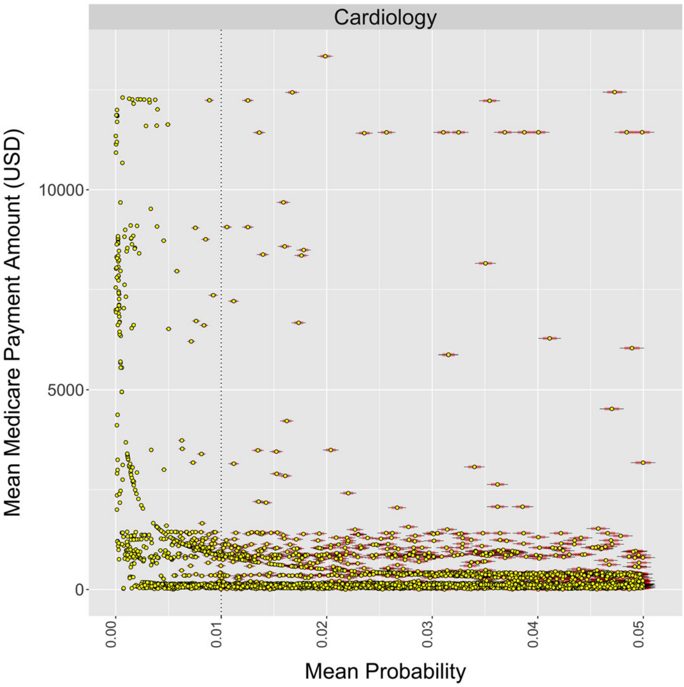

Fig. 9 Cardiology credible intervals

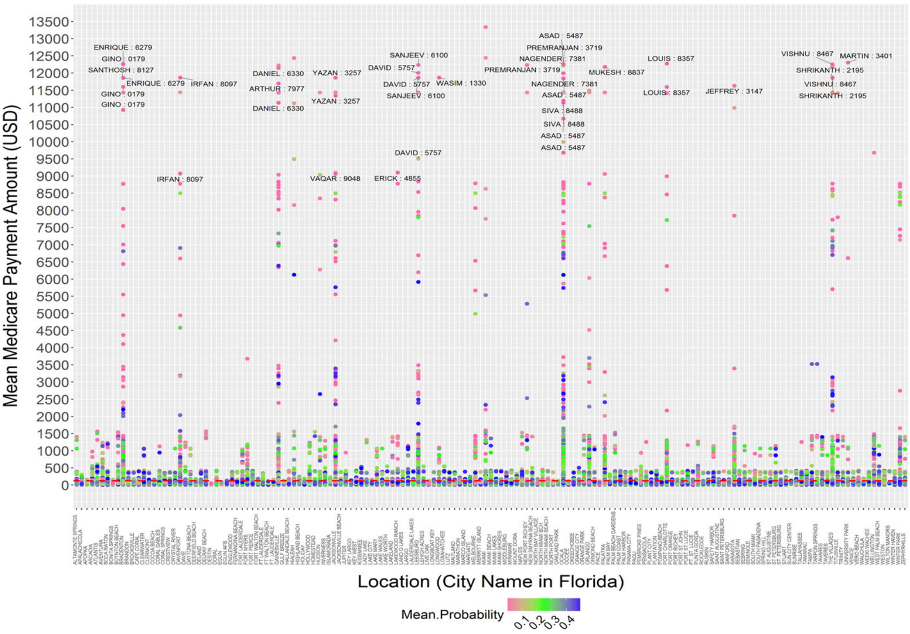

Fig. 10 Cardiology possible fraudulent activities by location

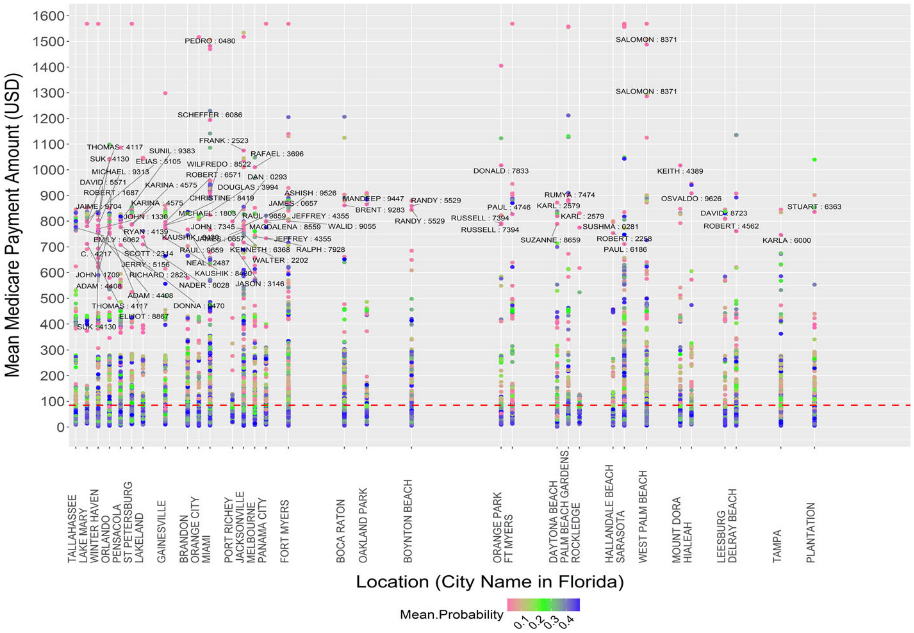

Fig. 11 Ophthalmology possible fraudulent activities by location

known fraudulent activities. Continued assessment and validation are left for future research. We have summarized the advantages of our model and demonstrated the ability of our outlier detection method to find possibly fraudulent behaviors. Moreover, our method is shown to be flexible in providing both an automatic way to model and account for multiple input variables and produce meaningful outputs consisting of probability distributions per point value (in this case per payment).

9 Conclusion

Medicare fraud continues to be a burden on the US healthcare systems requiring novel and pragmatic solutions. The recent public availability of Medicare data is a boon for continued research into methods to detect, investigate, and reduce healthcare fraud. There are many studies demonstrating possible ways of detecting fraudulent behaviors in healthcare, but very few focusing on Medicare or using multiple input variables to identify and assess anomalous activities.

Given this need and the availability of data, outlier detection is an important technique in finding anomalous activities, or behaviors, that could indicate possible fraud. The purpose of this study is to introduce a new, general methodology for multivariate outlier detection to discover possibly fraudulent activities in Medicare claims payment data. Our proposed method is a twofold approach that considers information from multiple input variables and provides meaningful outlier probabilities per completed payment. In order to do this, we first create a multivariate model using MARS to produce studentized residuals.

Then, we create a general, fully Bayesian probability model to detect anomalous values using the Stan probabilistic programming language. The residuals from the regression model are inputs into the probability model, which produces probabilities indicating possible anomalous activities. To thoroughly assess our method, we provide univariate and multivariate comparisons against other, common outlier detection methods. We then apply our method to the Medicare claims dataset to demonstrate its efficacy and flexibility.

Our analysis indicates that our outlier detection model performs favorably to other detection methods, while providing more meaningful information for further investigation. Unlike those methods, ours provides additional information such as credible intervals. We do not require parameters to designate the number of neighbors or clusters, or arbitrary thresholds based on non-meaningful distance measures. Additionally, our method was shown to be more robust to data distributions and predictor influence versus Mahalanobis distance and k-means clustering. Finally, we show how our method behaves using 7 different Medicare specialties and ways to view and interpret the outlier probabilities. The resulting probability distributions, per claim, coupled with the credible intervals, provide more information and give greater confidence in flagging and investigating illegitimate Medicare claims. To further bolster the usefulness of our approach, we provide examples incorporating two providers (Cardiology and Ophthalmology) under investigation for Medicare fraud.

Throughout our paper, we have discussed possible avenues for future work. Ongoing research and future work will involve including more medical specialties. Additionally, expanding the Medicare dataset beyond Florida and integrating additional datasets, such as referrals, could be assessed in order to detect a wider range of fraudulent activities. As previously mentioned, additional validation and testing against real-world cases should be performed to verify and enhance our current detection methodology. Lastly, our method is general and can be applied to detect anomalies over multivariate data in other domains, such as health monitoring.

Acknowledgements The authors would like to thank the anonymous reviewers and the Associate Editor for the constructive evaluation of this paper and also the various members of the Data Mining and Machine Learning Laboratory, Florida Atlantic University, for the assistance with the reviews. Also, we acknowledge partial support by the NSF (CNS-1427536). Opinions, findings, conclusions, or recommendations in this paper are the authors' and do not reflect the views of the NSF.

Compliance with ethical standards

Conflict of interest Richard A. Bauder declares that he has no conflict of interest. Taghi M. Khoshgoftaar declares that he has no conflict of interest.

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

Appendix 1: Additional information on Bayesian inference and probabilistic programming

The use probabilistic programming, as a general inference technique, enables us to easily create a model of some event or characteristic, such as the detection of outliers, and perform probabilistic reasoning for prediction, infer causes from past events, and learn from the past to improve predictions. At its core, probabilistic programming is probabilistic reasoning, where the probabilistic model is expressed using a programming language. This is, in part, because the probabilistic program is interpreted as the distribution from which one can use tools to ask questions about the distribution. Additionally, these modeling languages incorporate random events as primitives and, as mentioned, their runtime environment handles inference. There are other representation-type languages available, such as Bayesian Belief Networks and hidden Markov models. However, these methods are simply simulations, which is not machine learning. A probabilistic program is akin to a simulation that you can run and analyze.

Even though Bayesian inference updates the posterior probabilities, or beliefs, based on any prior assumptions and the evidence, the updated beliefs may not necessarily agree with our prior assumptions and, as in the real world, the evidence tends to bolster, or overtake, any prior assumptions we may have had concerning some event. In that case, we can assume non-informative, or vague, prior beliefs with little to no prior knowledge. We also have the ability to incorporate stronger beliefs and assumptions based on prior knowledge, such as expert opinions, thus improving the model fit. Other benefits of Bayesian methods include more interpretable results, the incorporation of subjective inputs such as medical knowledge, and the quantification of uncertainties. Furthermore, Bayesian techniques provide credible intervals for the different parameters in the model. The credible intervals show that a value or parameter has, for instance, an 80 or 95% probability of being within the interval bands (note these are the default intervals exported by Stan). This is much easier to interpret than the traditional confidence intervals, which indicate that if an experiment is repeated many times, the values will be within this interval 80 or 95% of the time. Additionally, as with other means to assess confidence in results, the credible intervals around the results are reliable given that the model is true.

Appendix 2: Probability model code and details

Below is our general outlier detection Stan probability model, with comments, showing inputs, unknown variables, distributions, and generated outputs. Note the Stan language is mostly vectorized, but there are instances where vectorization is not yet implemented. For instance, in the Stan probability model code, the prior distributions are vectorized functions whereas the likelihood distribution is not hence the need for a *for* loop. Even so, both methods result in their requisite distributions.

```
data{
    int<lower=0> N;
    vector[N] value;        // input values
    vector[N] check_value;  // check values
}

parameters{
    real mean_value;
    real stdev_value;
    real nu;                // deg of freedom
}

model{
    // distribution of mean
    mean_value ~ normal(100, 100);

    // distribution of standard deviation
    stdev_value ~ normal(100, 100);

    // degrees of freedom, lower bound of zero
    nu ~ cauchy(7, 5) T[0.0,];

    // Student's t-distribution for outliers
    for(i in 1:N)
      value[i] ~ student_t(nu, mean_value,
                           stdev_value);
}
generated quantities{
    // cumulative dist of probabilities
    vector[N] cdf_prob;

    // inverse of cdf
    vector[N] ccdf_prob;

    // final outlier probabilities
    vector[N] prob;

    for(i in 1:N){
      cdf_prob[i] = student_t_cdf(check_value[i],
                                  nu, mean_value,
                                  stdev_value);
      ccdf_prob[i] = 1 - cdf_prob[i];
      prob[i] = 2*(cdf_prob[i]*ccdf_prob[i]);
    }
}
```

In the case of the prior distributions given by $mean\_value \sim normal(100, 100)$; and $stdev\_value \sim normal(100, 100)$; in the Stan model code, both the mean and standard deviation are prior distributions declared as normal distributions with mean 100 and standard deviation 100. The use of 100 is not set in stone, but more a starting point for the distributions used to generate the likelihood distribution for the student's *t* distribution. These static parameters are typically chosen by the researcher as assumptions on the prior distributions, or prior knowledge of the data.

References

Aggarwal, C.C.: Data Mining: The Textbook. Springer, Berlin (2015). google-Books-ID: cfNICAAAQBAJ

Aggarwal, C.C.: Outlier analysis. In: Data Mining. Springer, Berlin, pp. 237–263 (2015)

Bauder, R., Khoshgoftaar, T.M., Seliya, N.: A survey on the state of healthcare upcoding fraud analysis and detection. Health Serv. Outcomes Res. Methodol. **17**(1), 31–55 (2017). doi:10.1007/s10742-016-0154-8. [Online]

Ben-Gal, I.: Outlier detection. In: Data Mining and Knowledge Discovery Handbook. Springer, Berlin, pp. 131–146 (2005)

Berwick, D.M., Hackbarth, A.D.: liminating waste in us health care. JAMA **307**(14), 1513–1516 (2012). doi:10.1001/jama.2012.362

Box, G.E., Tiao, G.C.: Bayesian Inference in Statistical Analysis, vol. 40. Wiley, New York (2011)

Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. In: ACM sigmod record, vol. **29**, no. 2, pp. 93–104. ACM (2000)

Brooks, S., Gelman, A., Jones, G., Meng, X.-L.: Handbook of Markov Chain Monte Carlo. CRC press, Boca Raton (2011)

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., Riddell, A.: Stan: a probabilistic programming language. J. Stat. Softw. **20** (2016)

Centers for Medicare and Medicaid Services Frequently Asked Questions. https://questions.cms.gov/

Centers for Medicare and Medicaid Services: HCPCS General Information. https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/

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

Chaloner, K., Brant, R.: A Bayesian approach to outlier detection and residual analysis. Biometrika, **75**(4), 651–659 (1988). http://biomet.oxfordjournals.org/content/75/4/651.abstract

Compute Mahalanobis Distance and flag multivariate outliers (Sep 2016). http://www-01.ibm.com/support/docview.wss?uid=swg21480128

Cousineau, D., Chartier, S.: Outliers detection and treatment: a review. Int. J. Psychol. Res. **3**(1), 58–67 (2015)

DARPA probabilistic programming for advancing machine learning (PPAML) (2016). http://www.darpa.mil/program/probabilistic-programming-for-advancing-machine-Learning

Das, M.K., Gogoi, B.: Usage of graphical displays to detect outlying observations in linear regression. Indian J. Appl. Res. **5**(5) (2016)

Davenport, K.: Mahalanobis Distance and Outliers (2013). http://kldavenport.com/mahalanobis-distance-and-outliers/

Davidson-Pilon, C.: Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference. Addison-Wesley Professional, Boston (2015)

Dobson, A.J., Barnett, A.: An Introduction to Generalized Linear Models. CRC press, Boca Raton (2008)

Ekina, T., Leva, F., Ruggeri, F., Soyer, R.: Application of Bayesian methods in detection of healthcare fraud. Chem. Eng. Trans. **33** (2013)

Friedman, J.H.: Multivariate adaptive regression splines. Ann. Stat. **19**(1), 1–67 (1991)

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis, vol. 2. CRC Press, Boca Raton (2014). doi:10.1080/01621459.2014.963405

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

Greenburg, J.: Medicare fraud rate is 8 to 10 percent, says Roskam of Illinois (2013). http://www.politifact.com/truth-o-meter/statements/2013/jun/17/peter-roskam/rep-roskam-says-medicare-fraud-rate-8-10-percent/

Grubbs, F.E.: Sample criteria for testing outlying observations. Ann. Math. Stat. 27–58 (1950)

Grubbs, F.E.: Procedures for Detecting Outlying Observations in Samples, vol. **11**(1), 1–21 (1969). http://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657

Guo, J., Gabry, J., Goodrich, B., Lee, D., Sakrejda, K., Sklyar, O., Oehlschlaegel-Akiyoshi, J., Wickham, H., de Guzman, J., Fletcher, J., Heller, T., Niebler, E., the R Core Team: RSTAN: R Interface to Stan (2016). https://cran.r-project.org/web/packages/rstan/index.html

Heathcare.gov glossary (2017). https://www.healthcare.gov/glossary/

Hiers, F.: Cardiologist plagued by legal woes files for Chapter 11 bankruptcy protection. http://www.ocala.com/article/20160422/ARTICLES/160429933

How Growth of Elderly Population in US Compares With Other Countries (2013). http://www.pbs.org/newshour/rundown/how-growth-of-elderly-population-in-us-compares-with-other-countries/

Hu, J., Wang, F., Sun, J., Sorrentino, R., Ebadollahi, S.: A healthcare utilization analysis framework for hot spotting and contextual anomaly detection. In: AMIA (2012)

Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., Arab, M.: Using data mining to detect health care fraud and abuse: a review of literature. Glob. J. Health Sci. **7**(1), 194–202 (2014)

Komsta, L.: Package ‘outliers (2015). https://cran.r-project.org/web/packages/outliers/outliers.pdf

Korkmaz, S., Goksuluk, D., Zararsiz, G.: MVN: An R package for assessing multivariate normality. R J. **6**(2), 151–162 (2014). http://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C.: CARET: Classification and Regression Training, 2016, R package version 6.0-68. https://CRAN.R-project.org/package=caret

Li, J., Huang, K.-Y., Jin, J., Shi, J.: A survey on statistical methods for health care fraud detection. Health Care Manag. Sci. **11**(3), 275–287 (2008)

Local Outlier Factor (2016). https://turi.com/learn/userguide/anomaly_detection/local_outlier_factor.html

Matthews, S.: How the Aging Population Is Changing the Healthcare System (2013). http://www.everydayhealth.com/senior-health/aging-and-health/pressures-on-healthcare-from-booming-senior-population.aspx

Milborrow, S., Hastie, T., Tibshirani, R.: Earth: Multivariate Adaptive Regression Splines (2016). https://cran.r-project.org/web/packages/earth/index.html

Mínguez, R., Reguero, B.G., Luceno, A., Méndez, F.J.: Regression models for outlier identification (hurricanes and typhoons) in wave hindcast databases. J. Atmos. Ocean. Technol. **29**(2), 267–285 (2012). doi:10.1175/JTECH-D-11-00059.1

MINITAB 17 Statistical Software [Computer software] (2010). www.minitab.com

Morris, L.: Combating Fraud In Health Care: An Essential Component of Any Cost Containment Strategy (2009). http://content.healthaffairs.org/content/28/5/1351.full

Multivariate Adaptivie Regression Splines (2015). https://documents.software.dell.com/statistics/textbook/multivariate-adaptive-regression-splines

Munro, D.: Annual U.S. healthcare spending hits $3.8 trillion (2014). http://www.forbes.com/sites/danmunro/2014/02/02/annual-u-s-healthcare-spending-hits-3-8-trillion/

National Health Accounts by service type and funding source (2014). https://www.cms.gov/Research-Statistics-Data-and-systems/Statistics-Trends-and-reports/NationalHealthExpendData/index.html

National Plan & Provider Enumeration System (NPPES): National Provider Identifier. https://nppes.cms.hhs.gov/NPPES/

Nedret, B., Gulsen, K.: A comparison of multiple outlier detection methods for regression data. Commun. Stat. Simul. Comput. **37**(3), 521–545 (2008). doi:10.1080/03610910701812352

NIST/SEMATECH e-Handbook of Statistical Methods (Anderson-Darling) (2013). http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm

NIST/SEMATECH e-Handbook of Statistical Methods (Grubb’s test) (2013). http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

NOLO: Part B Medical Insurance: What You Pay (2010). http://www.nolo.com/legal-encyclopedia/part-b-medical-insurance-what-you-pay.html

Office of Inspector General: Exclusions Program. http://oig.hhs.gov/exclusions/index.asp

Pande, P.S., Neuman, R.P.: The Six Sigma Way. McGraw-Hill, New York (2000)

Profile of older Americans: 2015 (2015). http://www.aoa.acl.gov/Aging_Statistics/Profile/2015/

R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, (2017). https://www.R-project.org/

Robinson, D.: K-means clustering is not a free lunch (2015). http://varianceexplained.org/r/kmeans-free-lunch/

Rosenmai, P.: Using Mahalanobis Distance to Find Outliers (2013). http://eurekastatistics.com/using-mahalanobis-distance-to-find-outliers/

Ross, S.: A First Course in Probability, 6th edn. Prentice-Hall, Upper Saddle River (2002)

Savage, J.: A quick-start introduction to Stan for economists (2016). http://quantecon.org/notebooks.html

Stan Development Team, Stan Modeling Language Users Guide and Reference Manual

State Health Facts-Medicare (2015). http://kff.org/state-category/medicare/

Steinbusch, P.J., Oostenbrink, J.B., Zuurbier, J.J., Schaepkens, F.J.: The risk of upcoding in casemix systems: a comparative study. Health Policy **81**(2–3), 289–299 (2007)

Stevens, J.P.: Outliers and influential data points in regression analysis. Psychol. Bull. **95**(2), 334 (1984)

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

Tallarida, R.J., Murray, R.B.: Chi-square test. In: Manual of Pharmacologic Calculations. Springer, Berlin, pp. 140–142 (1987)

The facts about rising health care costs (2015). http://www.aetna.com/health-reform-connection/aetnas-vision/facts-about-costs.html

Thornton, D., Capelleveen, G., Poel, M., Hillegersberg, J., Müller, R.M.: Outlier-based health insurance fraud detection for us medicaid data (2014)

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

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

Understanding Medicare-allowed amounts (2016). https://secure.wpsic.com/sales-materials/files/28807-medicare-approved-amounts-tip-sheet.pdf

US Medicaid Program (2016). https://www.medicaid.gov

US Medicare Program (2016). https://www.medicare.gov

Weaver, J., Chang, D.: South Florida ophthalmologist emerges as Medicare's top-paid physician. http://www.miamiherald.com/news/local/community/miami-dade/article1962581.html

Williamson, D.F., Parker, R.A., Kendrick, J.S.: The box plot: a simple visual method to interpret data. Ann. Intern. Med. **110**(11), 916–921 (1989). doi:10.7326/0003-4819-110-11-916

Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, Burlington (2005)

Zhang, J.: Advancements of outlier detection: a survey. ICST Trans. Scalable Inf. Syst. **13**(1), 1–26 (2013)