2016 15th IEEE International Conference on Machine Learning and Applications

A Probabilistic Programming Approach for Outlier Detection in Healthcare Claims

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

***Abstract*—**

Healthcare is an integral component in people's lives, especially for the rising elderly population. Medicare is one such healthcare program that provides for the needs of the elderly. It is imperative that these healthcare programs are affordable, but this is not always the case. Out of the many possible factors for the rising cost of healthcare, claims fraud is a major contributor, but its impact can be lessened through effective fraud detection. We propose a general outlier detection model, based on Bayesian inference, using probabilistic programming. Our model provides probability distributions rather than just point values, as with most common outlier detection methods. Credible intervals are also generated to further enhance confidence that the detected outliers should in fact be considered outliers. Two case studies are presented demonstrating our model's effectiveness in detecting outliers. The first case study uses temperature data in order to provide a clear comparison of several outlier detection techniques. The second case study uses a Medicare dataset to showcase our proposed outlier detection model. Our results show that the successful detection of outliers, which indicate possible fraudulent activities, can provide effective and meaningful results for further investigation within medical specialties or by using real-world, medical provider fraud investigation cases.

***Keywords*—Fraud Detection, Outlier Detection, Healthcare Fraud, Bayesian Inference, Probabilistic Programming**

I. Introduction

Healthcare is an essential, yet costly, part of most people's lives. The costs associated with most healthcare plans can cripple individuals and families leading to increased stress and potentially life altering choices. Even given such social prominence, healthcare costs continue to rise with fraud, waste, and abuse (FWA) reduction efforts doing little to diminish these costs [7]. The problem is further exacerbated by the rising elderly population. One of the largest programs to provide insurance to the growing elderly population in the US is Medicare.

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 [9]. To put the value of Medicare in context, the National Health Expenditures 2013 Highlights [6], released by the Centers for Medicare and Medicaid (CMS) [1], indicate the US healthcare spending in 2013 increased 3.6%, from 2012, to reach $2.9 trillion. Out of this, Medicare spending was about 20% of the national healthcare spending at nearly $587 billion, which is an increase of 3.4% over 2012. It is estimated that for Medicare alone, recovery of 10% to 15% of expenses is possible through effective and efficient fraud detection [27]. With these large dollar values increasing annually, and limited reductions due to FWA efforts, new methods to detect and flag possible FWA events are needed.

In this paper, our contribution is in creating a generalizable, fully Bayesian probability model to detect anomalous values using probabilistic programming. In relation to our study, an anomalous value may indicate a medical claims fraud case. 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. 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. Our model is shown to provide meaningful information beyond common outlier detection methods and produce relevant outlier results for further investigation. In order to demonstrate the capabilities and generalizable nature of our approach, we provide two case studies. The first case study involves a temperature dataset, with added outliers, to clearly demonstrate the capabilities and results of our method versus several common outlier techniques. The second case study uses real-world Medicare data [4] to showcase our model, which includes highlighting the value of credible intervals in detecting claims anomalies, as well as provide a real fraud-related example [31] as a form of validation. Since fraud cases are relatively rare and our method is general, there is no focus on any particular type of fraud [28], such as upcoding [11] or self-referrals. Moreover, this limited number of known Medicare fraud cases, or ground truth, makes meaningful comparative outlier detection analysis difficult with very few examples of actual outliers to validate each model against, thus the inclusion of the first case study. To the best of our knowledge, there are no related studies that build and use a probabilistic programming model to detect outliers in Medicare (or healthcare) data.

The rest of the paper is organized as follows. Section II discusses works related to the current research in this domain. In Section III, we review several outlier detection methods. Section IV explains the Bayesian probability model and probabilistic programming. Sections V and VI detail our two case studies, to include data, methods, discussion, and results. Finally, Section VII outlines our conclusions and ideas for future work.

978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.28

II. Related Works

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 [26], [24], [29]. Bayesian techniques are less common but still well represented in the literature, especially Naïve Bayes. Tomar et al. [29], and Kolce and Frasheri [25] consider both Naïve Bayes and Bayesian Belief Networks as data mining approaches for specific problems in healthcare. 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 [33]. The preponderance of references for full Bayesian techniques, in healthcare, are focused more on disease diagnosis or prognosis [25] or improving detection of physiological readings [17]. Thus, many full Bayesian methods are not directly related to healthcare FWA. This may be because Bayesian models do not give completely accurate results, but rather a distribution of results that encompass uncertainty. This is more akin to real-world scenarios, which can be more beneficial when dealing with the detection of outliers, such as with healthcare fraud.

One of the more related works, incorporating both healthcare fraud and Bayesian methods, is by Ekin et al. [20], which is also referenced in [24]. 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. The data is organized in a matrix where rows are providers and columns are beneficiaries. Each instance, or row, in the matrix is then a row-column cluster, or co-cluster. 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 the probabilities of latent variables. These posterior distributions are used to flag potential fraudulent activities via unusual cluster memberships. The authors test their algorithm on simulated data, thus do not provide any comparative analysis or tangible results for their Bayesian approach to healthcare fraud detection.

Wang and Luo [30] incorporate a probabilistic programming model using Stan [16] to create an improved Beta regression model. Mixed effect Beta regression models are common when researchers analyze clinical trial results, to model the co-variate effects on proportional or percentage responses through a generalized linear model. These clinical trials, whether longitudinal percentage or proportional data, contain values between zero and one. The authors propose a mixed effects model using a Beta rectangular distribution and augment it with the probabilities for the closed interval of zero and one using a generalized Bayesian method. They create a one-augmented Beta rectangular regression (OABR) model with Inverse-Gamma and Uniform weak prior distributions. Posterior distribution samples for each unknown parameter are obtained using Hamiltonian Monte Carlo and No-U-Turn Sampler, which are all implemented in Stan. They use Parkinson's disease long-term study data, specifically the EuroQol vertical visual analog scale (EQ-VAS) metric. The authors compare their OABR model against the one-inflated Beta regression model (BEOI) with results suggesting their model generates more consistent outcomes when data are simulated using either their OABR model or the BEOI model, with both the original and contaminated datasets.

Both Ekin et al. [20] and Wang and Luo [30] discuss fully Bayesian techniques using healthcare-related datasets. The former is a Bayesian approach to healthcare fraud using Bayesian co-clustering, but is limited in scope considering only simulated data and relying on group memberships to assess possible fraud events. The latter incorporates a probabilistic programming model, using Stan, for healthcare clinical trial results, but does not address FWA concerns or emphasize the detection of anomalous values. For this paper, we use the 2012 - 2014 Medicare data to demonstrate the efficacy of our Bayesian outlier detection model, incorporating a probabilistic programming approach. The work presented herein is novel in its development and use of a probabilistic programming language to create a generalizable outlier detection model for healthcare fraud, waste, and abuse.

III. Outlier Detection Methods

We briefly discuss several commonly used outlier detection methods for which we provide a comparative analysis in Section V. While this paper does not provide a comprehensive discussion on outlier detection methods, we refer the reader to [10], [34]. An outlier is simply some value that lies outside of the main group that it is a part of. 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. Another popular, and related method, is to take any point more than *t* standard deviations from the mean as an outlier. The problem with this method is that both the mean and standard deviation are quite sensitive to the presence of outliers in the dataset.

The standard boxplot rule [32] uses the quartiles of a dataset, thus does not require knowledge of all the data to create specified thresholds and is less subject to the effects of outliers on that dataset. More specifically, the upper and lower bounds, which indicate possible outliers, are calculated via the interquartile range, as seen in Equation 1. The value *c* is typically 1.5 or 3.0 to indicate outliers.

$$\begin{aligned} Threshold = (Q_1 - c \times (Q_3 - Q_1), \\ Q_3 + c \times (Q_3 - Q_1)) \end{aligned} \tag{1}$$

Instead of relying on the mean and standard deviation, the Hampel indicator [12] replaces the mean with median and the standard deviation with the Median Absolute Deviation (MAD) scale. MAD is generally more effective than the other threshold-based methods, because it is less prone to the effects of outliers in the dataset. Equation 2 shows how to create lower and upper thresholds based on the median and MAD scale.

$$\begin{aligned} MAD &= median(|x_i - median(x)|) \\ Threshold &= (median(x) - c \times MAD, \\ &\qquad median(x) + c \times MAD) \end{aligned} \tag{2}$$

The methods discussed thus far are threshold-based ways to detect outliers in a dataset. These methods tend to miss real outlier values, especially in datasets with a lot of variability. Another effective way to detect outliers, and better handle this variability, is to use density- or distance-based techniques. One such method is known as Local Outlier Factor (LOF) [14]. 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. An average ratio of the points reachability and its *k*-nearest neighbors' local densities is used as the LOF score. 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.

The last outlier method for our comparison is the Grubbs' test, based on Grubbs' procedures for detecting outliers [23]. 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 normality, which may not be applicable to some datasets requiring data transformation or using a different test for outliers [5].

IV. Probability Model

A. Probabilistic Programming and Bayesian Inference

In this study, we use probabilistic programming [22], [15], [19], [16] to detect outliers in healthcare-related 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 probabilistic programming initiative through the Defense Advanced Research Projects Agency [8], which began in 2013.

Probabilistic programming allows for general statistical inference via the probabilistic language's runtime environment which, essentially, runs the program both forward and backward. The forward run goes from causes to effects (i.e. data) with the backward run going from data to the causes [18]. At its core, probabilistic programming is probabilistic reasoning, where the probabilistic model is expressed using a programming language [19]. There are other representation-type languages available, such as Bayesian Belief Networks and hidden Markov models. However, these methods are simply simulations, whereas a probabilistic program is akin to a simulation that you can run and analyze [19].

In our study, we use the power of probabilistic programming to perform full Bayesian inference. We provide a summary of some of the main points of Bayesian inference, with more detailed information available in [13]. Bayesian inference provides a way of combining new evidence with prior beliefs, or assumptions, through the application of Bayes' rule, defined in Equation 3.

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)} \tag{3}$$

In Bayes' rule, P(A) is the prior belief in event A (previous assumptions), P(X) is the prior probability of the evidence, P(X|A) is the likelihood of evidence X given event A, and P(A|X) is the posterior probability, which is the updated belief based on the evidence. This equation basically 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.

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 an 80% or 95% probability of being within the interval bands. 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.

B. Outlier Detection Model

To create our outlier detection probability model, we use the probabilistic programming language known as Stan [16]. 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. Additionally, for model completeness and convergence, we run multiple (2 or 4 based on the dataset size and modeling performance, where 4 is the default in Stan [16]) Markov chains [15] and ensure the split-$\hat{R}$ multi-chain diagnostic for all parameters is smaller than 1.1 [22]. We are using the rstan[1] implementation of the Stan probabilistic programming language. Below is our general outlier detection Stan model showing inputs, unknown variables, distributions, and generated outputs.

```
data{
    int<lower=0> N;
    int<lower=0> I;
    vector<lower=0>[N] value;
    vector[I] value_thresholds;
}

parameters{
```

[1] http://mc-stan.org/interfaces/rstan

```
    real mean_value;
    real stdev_value;
}

model{
    mean_value ~ normal(1000, 1000);
    stdev_value ~ normal(1000, 1000);

    for(i in 1:N)
        value[i] ~ normal(mean_value, stdev_
            value) T[0.0,];
}

generated quantities{
    vector[I] cdf_prob;
    vector[I] ccdf_prob;
    vector[I] prob;

    for(i in 1:I) {
        cdf_prob[i] <- exp(normal_cdf_log(
            value_thresholds[i], mean_value,
            stdev_value));
        ccdf_prob[i] <- exp(normal_ccdf_log(
            value_thresholds[i], mean_value,
            stdev_value));
        prob[i] <- 2 * (cdf_prob[i] * ccdf_
            prob[i]);
    }
}
```

This study is not a tutorial on Stan, but the interested reader can find additional information in [16]. However, in order to get a conceptual understanding of the outlier detection model, we provide details on specific parts of the above Stan model. Our current outlier detection model is a univariate model, where a multivariate model and/or establishing relationships between two or more variables is left for future work.

We declare four input variables that correspond to the actual inputs including the vector for the population (or sample) from which to assess outliers, the thresholds or values to check and determine outliers, and the lengths of each of these vectors. The unknown model parameters for Stan to estimate are the mean and standard deviation of the sample space (from which to detect outliers). For the detection of outliers, we assume a normal distribution for the probabilities, using mean and standard deviation, as defined in Equation 4 showing the probability density and standard Gaussian distribution.

$$f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \\ X \sim Normal(\mu,\sigma^2) \tag{4}$$

We define non-informative, or vague, priors of normal distributions for both the mean and standard deviation. From these, the likelihood of the value is defined as a normal 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 driven by the evidence. Finally, we declare the quantities other than the simulated parameters to be generated. In this case, we simply generate cumulative normal distribution functions (cdf) for being both greater than and less than some value, where the normal cdf is defined in Equation 5 with variables from the generated quantities Stan model block.

$$\int_{-\infty}^{value\_threshold} Normal(x|\mu,\sigma^2)dx \tag{5}$$

We compute the probability of observing a more extreme value, stored in the 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 "bell curve", thus unlikely to be an outlier. In contrast, a probability of 4% would indicate only a 4% 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 a normal distribution.

V. Case Study: Comparison of Outlier Detection Methods

A. Temperature Data

For this case study, we use temperature data to compare various outlier detection techniques, i.e. those presented in Section III and our probabilistic programming model. We chose temperature because it is a well-known variable with easily understood normal ranges, for both weather experts and novices alike, allowing for succinct comparative analysis. The data, measured in Fahrenheit (°F), is from Weather Underground[2] for March and April of 2016 in Florida. There are 13 artificially edited points that are meant to portray clear and distinct outliers. The range for the daily temperature averages is 66°F to 79.4°F, with an average of 72.61°F. The edited outlier points are all above 80°F and below 62°F, with several of the points being boundary values that are close to normal ones yet still considered outliers.

B. Methodology

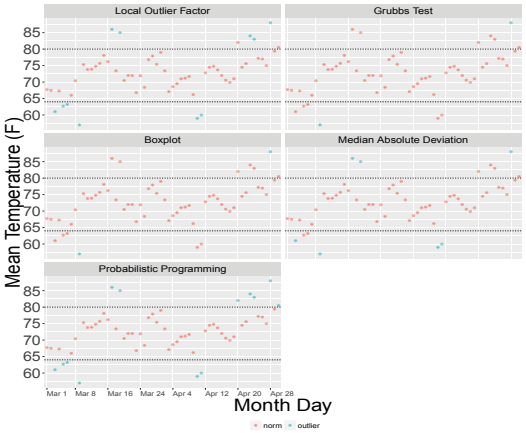
As is typical with outlier detection methods, the input data for the Stan model has identical vectors for the *value* and *value_thresholds* variables. This implies that we are looking for any outliers in the full population (only two months in this case study). Our model is configured to run 4,000 iterations with 4 Markov chains (the Stan default value) on this vector of temperatures. Using the same vector of temperatures, thresholds are defined based on the Median Absolute Deviation with a factor of 3 to indicate an outlier. For LOF, several values of *k* were tried, including 5, 10, 25, and 40, in order to optimize the discovered outliers. We ended up using 29 neighbors, which is half of the length of the temperature vector. A two-sided Grubbs' test was run on the entire temperature dataset. Lastly, the boxplot rule was applied with the standard value of 1.5 used to indicate outliers. We did not perform the boxplot rule or Grubbs' test iteratively, thus only two outlier points are found. The incorporation of iteration would increase the number of outliers discovered, but it also complicates the search for outliers by adding complexity in finding the step or bin sizes required for optimal outlier detection.

[2] www.wunderground.com

C. Results and Discussion

Each of the outlier detection methods are capable of flagging obvious outliers, such as 57°F and 88°F. Grubbs' test and the boxplot rule both marked the same values as outliers, which coincide with the most extreme temperature values. LOF and MAD have comparable outlier results, but neither correctly capture the outliers on the boundaries. Our probability model, however, successfully flagged all 13 outlier values to include the boundary cases. Figure 1 shows each of the outlier detection methods and those points marked as outliers. The normal temperature values are surrounded by black boundary lines at 62°F and 82°F.

Fig. 1: Comparison of Outlier Detection Methods



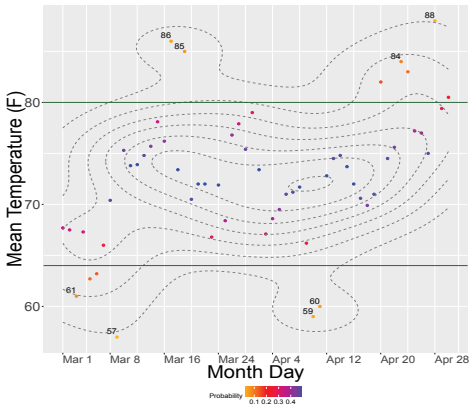
In comparison to our probability model, the other outlier detection methods provide point values or thresholds for possible outliers, but do not provide additional information like the probability of a value actually being an outlier. Even a more sophisticated method like LOF does not provide this, but rather gives a score which is subject to interpretation as to what constitutes a normal point or an outlier. Our model for outlier detection, which is full Bayesian inference, provides a distribution of probabilities per value providing information to help discern true outliers. In Figure 2, we depict the same temperature values, with mean probabilities per value, as well as contour lines indicating 5% probabilities with the central contours being 40% and 50% probabilities. This additional information shows us that temperatures, such as 85°F and 86°F, have low probabilities indicating very little chance of having more extreme values, thus should be considered outliers. The same holds for lower temperature values like 59°F and 57°F. In this case study, our model outperforms the other outlier methods while providing valuable information in determining what is or is not a legitimate outlier.

VI. Case Study: Probabilistic Outlier Detection for Medicare Payments

A. Medicare Data

The data that the Centers for Medicare and Medicaid Services [2] has released, at the point of this publication, is for calendar years 2012, 2013, and 2014. We use the Physician and Other Supplier Data 2012 - 2014 dataset, which describes

Fig. 2: Temperature Probabilities



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 only (as opposed to larger facilities, such as hospitals). Due to 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 [21]. In order to accurately identify providers, each physician is given a National Provider Identifier (NPI) [3]. Table I summarizes both the complete and Florida-only Medicare datasets.

TABLE I: 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	6,741
Florida	1,197,238	26	48,230	2,922

B. Methodology

For this case study, we focused on the dermatology and optometry provider specialties. We chose these, in part, due to the demographic of Florida and prevalence of certain specialties. Additionally, we used the claims data of an ophthalmology provider, which is different than optometry, from a known fraud case [31] to further validate our model, as well as demonstrate the flexibility in discovering outliers. Specifically, we use the Average Medicare Payment Amount and Line Service Count variables found in the Medicare dataset for our outlier detection model. For each of the provider specialties, we bin the data based on the total number of procedures performed. Each provider has a unique distribution for a given specialty, which could represent a different payment distribution based on the number of procedures. This binning process helps account for some of these variations in claims payments. We also incorporate several other variables that provide information on the providers, procedures performed, and locations in order to facilitate further investigation into the outliers as possible fraudulent activities. Table II lists and describes each of the pertinent variables.

TABLE II: Description of Medicare Variables

Variable Name	Description
Provider Type	Medical provider's specialty, e.g. Cardiology
HCPCS Code	Code for specific medical service furnished by the provider
NPI	Unique provider identification number
First Name	Provider's first name
Last Name	Provider's last name
City	Provider's office city
Line Service Count	Number of services provided / procedures performed
Avg. Medicare Payment Amount	Amount Medicare paid the provider for services performed

Moreover, this lack of known fraudulent cases, or ground truth, makes meaningful comparative outlier detection analysis difficult with very few examples of actual outliers to validate each model against, thus the inclusion of the first case study.

With optometry and dermatology, we took a subset of the binned procedures performed in order to clearly demonstrate and depict our model's outlier detection capabilities. For optometry, we include bins (360, 390] and (390, 420], with bins (1254, 1287] and (1287, 1320] used for dermatology. For these specialties, we run our outlier detection model in the same way as was done for the temperature data in Section V. Each of the binned medicare payment datasets were considered as the entire population, or sample, from which to assess possible payment outliers. The Stan model was run with 4,000 iterations and 2 chains (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 with a split-$\hat{R}$ of less than 1.1 for each value, over all Markov chains.

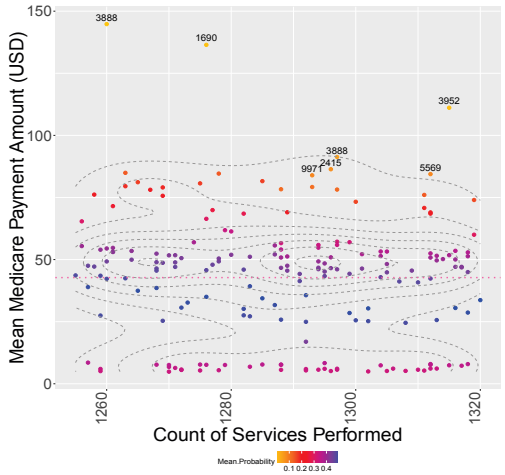
For the ophthalmology specialty, we searched for known fraud cases online in news reports. After searching for healthcare fraud in Florida, we discovered an ophthalmologist under criminal investigation for alleged excessive billing of Medicare [31]. It was reported that this provider's Medicare payments went toward reimbursements for injections of a costly drug, Lucentis, to treat patients with macular degeneration, a retinal disease. The Medicare payments from this provider, binned by the number of procedures performed, are compared against all ophthalmology payments in order to assess the probability of this specific provider's payments being outliers and possibly fraudulent. This is a different use of our outlier detection model, as we are comparing a smaller subset of values against the entire population to determine whether any of this subset should be considered outliers.

C. Discussion and Results

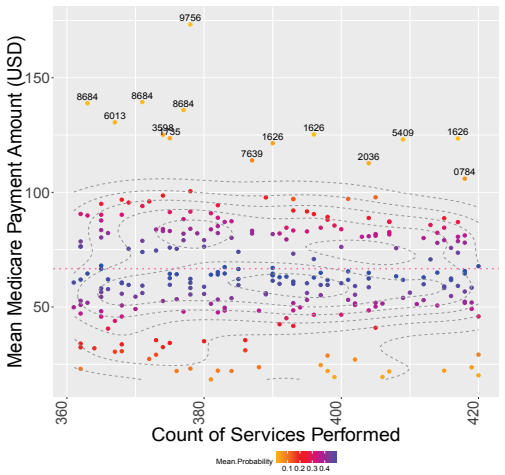
Figures 3a and 3b are Medicare payment probability plots indicating possible outliers for dermatology and optometry, respectively. We labeled any point below a 10% probability (indicating a possible outlier), represented by the pink horizontal line, and above the average value of all the data points. The threshold of 10% for the value labels is simply an aid to restrict possible outliers; however, all average probabilities are generated providing added information for identifying outliers. The use of points above the mean value simply focuses the outlier labels towards possible overpayment fraud activities, versus underpayment. The detected outliers for optometry are clearly delineated, given our labeling criteria, with the outlier values below a 10% probability depicting that most of the other values are less extreme. Also, some points below the mean are less than 10% probability, indicating possible outliers due to underpayment fraud activities, such as undercoding [11]. Dermatology shows clear outliers as well, but with more variance in the overall payment distribution. Even though the variance is higher, our model can still easily flag outliers and provide the probabilities to determine how possible outliers are affected by this uncertainty.

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. Figures 3c and 3d show the credible intervals for dermatology and optometry for the specified count of procedures performed. 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 10% probability threshold can be confidently flagged as either an outlier or normal. If the interval crosses over the 10% 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. Additionally, when comparing both dermatology and optometry credible intervals, the intervals are wider for dermatology 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.

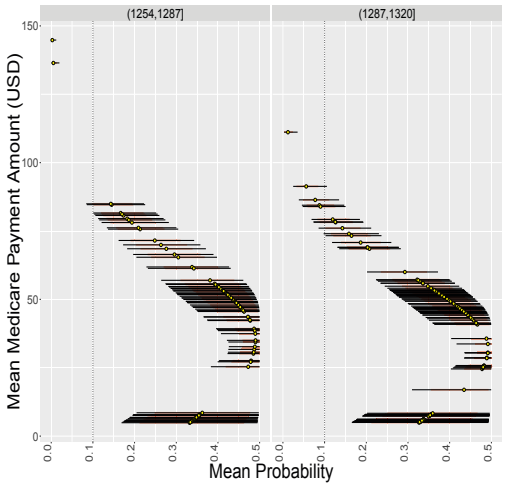
The outlier detection depicted in Figures 3a, 3b, 3c, and 3d flag outliers based on the entire subset of data, which is the same method employed in the temperature analysis in Section V. When looking at values for a specific ophthalmology provider, the input values for this provider are a subset of the total number of payment values in the ophthalmology specialty. Figure 3e shows the outlier probability plot for this provider versus all ophthalmology providers. The green points are all ophthalmology payments to provide a frame of reference when analyzing this particular provider's payment profile. The same threshold of 10% probability was applied, with the labeled values being those that are above the mean and less than the 10% probability. Even with this relatively small subset of data for the provider in question, a fairly large number of payment values are considered outliers indicating possible fraud. Figure 3f shows the credible intervals for the provider payment values. This again indicates the variance in the data, and probability intervals, from which the correct outliers can be confidently discerned. Most procedure count bins show fairly clear distinctions between values, with the exception of bin (128, 160]. This added uncertainty could be due to the limited ophthalmology payments within this range.



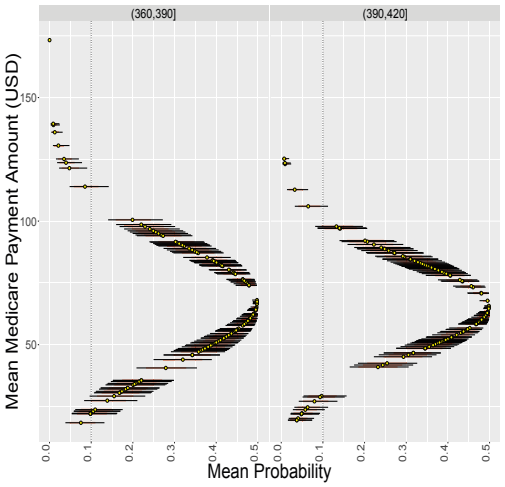
(a) Dermatology Outlier Probabilities



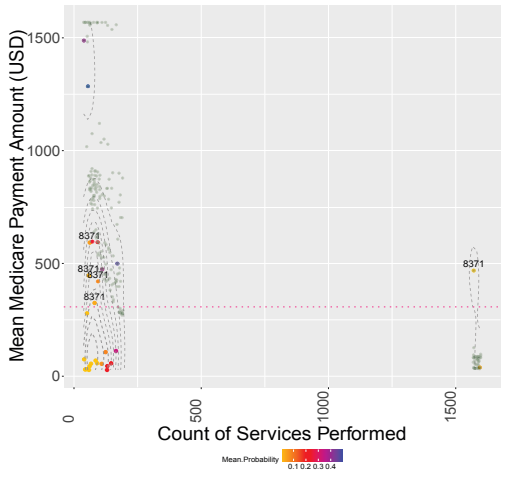
(b) Optometry Outlier Probabilities



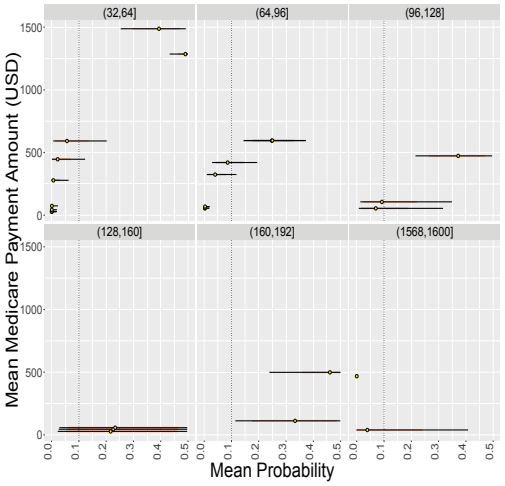
(c) Dermatology Credible Intervals



(d) Optometry Credible Intervals



(e) Outlier Probabilities for the Single Provider



(f) Credible Intervals for the Single Provider

Fig. 3: These plots depict the results of the second case study. The outlier probabilities of Medicare payments are shown indicating possibly fraudulent activities, as well as the credible intervals. The data is binned by the number of procedures performed, using a subset of bins for clarity.

VII. Conclusion

The detection of outliers is an important aspect in finding anomalous events in many varying domains. One such domain is healthcare where the misuse of medical insurance can lead to undesirable outcomes. The detection of outliers in claims payments can be used to detect possible fraud and provide a means to combat these wasteful and/or illegal activities. In this paper, we create a general, fully Bayesian probability model to detect anomalous values using the Stan probabilistic programming language. We demonstrate our outlier detection method in two case studies. The first case study uses temperature data, with added outliers, in order to compare several common outlier detection methods. Our outlier detection model was able to detect all outliers and provide probability distributions per value in order to further assess outlier validity. In the second case study, we incorporate real-world Medicare claims data to detect outliers. We show that, for dermatology and optometry, our model can detect possible fraudulent payments with meaningful probability information. Additionally, we provide an example incorporating a provider under investigation for Medicare fraud for further model validation. Future work should consider extending our model from a univariate to a multivariate model allowing for increased outlier detection capabilities by using the relationships between variables. Additional work could include a larger comparative study with Medicare data to evaluate the efficacy of different outlier detection models, including our proposed probability model.

Acknowledgment

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

References

[1] Centers for Medicare and Medicaid Services. [Online]. Available: https://www.cms.gov

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

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

[4] Physician and Other Supplier Data CY 2012 to 2014 - Centers for Medicare and Medicaid Services. [Online]. Available: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html

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

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

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

[8] "DARPA probabilistic programming for advancing machine learning (PPAML)," 2016. [Online]. Available: http://www.darpa.mil/program/probabilistic-programming-for-advancing-machine-Learning

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

[10] C. C. Aggarwal, "Outlier analysis," in *Data Mining*. Springer, 2015, pp. 237–263.

[11] R. Bauder, T. M. Khoshgoftaar, and N. Seliya, "A survey on the state of healthcare upcoding fraud analysis and detection," *Health Services and Outcomes Research Methodology*, pp. 1–25, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10742-016-0154-8

[12] I. Ben-Gal, "Outlier detection," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 131–146.

[13] G. E. Box and G. C. Tiao, *Bayesian inference in statistical analysis*. John Wiley & Sons, 2011, vol. 40.

[14] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[15] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[16] B. Carpenter, "Stan: A probabilistic programming language," *Journal of Statistical Software*, 2015.

[17] H. Chen, Y. Erol, E. Shen, and S. Russell, "Probabilistic model-based approach for heart beat detection," *arXiv preprint arXiv:1512.07931*, 2015.

[18] B. Cronin. What is probabilistic programming? [Online]. Available: radar.oreilly.com/2013/04/probabilistic-programming.html

[19] C. Davidson-Pilon, *Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference*. Massachusetts, USA: Addison-Wesley Professional, 2015.

[20] T. Ekina, F. Leva, F. Ruggeri, and R. Soyer, "Application of bayesian methods in detection of healthcare fraud," *Chemical Engineering Transaction*, vol. 33, 2013.

[21] T. H. J. K. F. Foundation, "State Health Facts - Medicare," 2015. [Online]. Available: http://kff.org/state-category/medicare/

[22] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Chapman & Hall/CRC Boca Raton, FL, USA, 2014, vol. 2.

[23] F. E. Grubbs, "Procedures for Detecting Outlying Observations in Samples," vol. 11, no. 1, Feb. 1969, pp. 1–21. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00401706.1969.10490657

[24] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab, "Using data mining to detect health care fraud and abuse: a review of literature." *Global journal of health science*, vol. 7, no. 1, pp. 194–202, 2014.

[25] E. Kolçe and N. Frasheri, "A literature review of data mining techniques used in healthcare databases," *ICT Innovations*, 2012.

[26] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," *Health care management science*, vol. 11, no. 3, pp. 275–287, 2008.

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

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

[29] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[30] J. Wang and S. Luo, "Augmented beta rectangular regression models: A bayesian perspective," *Biometrical Journal*, vol. 58, no. 1, pp. 206–221, 2016.

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

[32] D. F. Williamson, R. A. Parker, and J. S. Kendrick, "The Box Plot: A Simple Visual Method to Interpret Data," *Annals of Internal Medicine*, vol. 110, no. 11, pp. 916–921, Jun. 1989. [Online]. Available: http://dx.doi.org/10.7326/0003-4819-110-11-916

[33] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, Jul. 2005.

[34] J. Zhang, "Advancements of outlier detection: A survey," *ICST Transactions on Scalable Information Systems*, vol. 13, no. 1, pp. 1–26, 2013.