Overpayment Models for Medical Audits: Multiple Scenarios

Tahir Ekin, R. Muzaffer Musal and Lawrence V. Fulton [*]

01/11/2015

Abstract

Comprehensive auditing in Medicare programs is infeasible due to the large number of claims, therefore the use of statistical sampling and estimation methods is crucial. We introduce super-population models to understand the overpayment phenomena within the claims population. The zero and one inflated mixture based models can capture various overpayment patterns including the fully legitimate or fraudulent cases. We compare them with the existing models for symmetric and mixed payment populations that have different overpayment patterns. The distributional fit between the actual and estimated overpayments is assessed. We also provide comparisons of models with respect to their conformance with Centers for Medicare and Medicaid Services (CMS) guidelines. In addition to estimating the dollar amount of recovery, the proposed models can help the investigators to detect overpayment patterns.

Keywords : Audit, Simulation, Overpayment, Mixture, Zero-one inflated models

JEL Classification Codes : C, C1, C130, C150 , M420

[*] Ekin and Musal are at Department of Computer Information Systems and Quantitative Methods, Texas State University. Fulton is at Rawls College of Business, Texas Tech University. Dr. Ekin is the corresponding author with the contact information 601 University Dr. McCoy 411, San Marcos Texas 78666, 001 512 245 3297,t_e18@txstate.edu.

1 Introduction

Health care expenditures constitute a significant portion of the governmental budgets, especially in developed countries with high median age populations. For instance, in the United States, total health care related spending reached 17.9 percent of the GDP, 2.6 trillion USD, in 2010 (CMS [2010]). According to the U.S. federal agencies, every year three to ten percent of this spending is lost to Fraud, Waste and Abuse (FWA) (Shin et al. [2012]). Increasing budget deficits and deteriorating health care services over time have prompted the governmental organizations to launch more initiatives in order to control the health care spending and reduce the unnecessary payments. In addition to the direct cost implications, medical fraud diminishes the ability of the health care system to serve beneficiaries efficiently (Anderson and Hussey [2001]). This led to closer screening of medical claims via more investigations and the systematic use of statistical approaches. A comprehensive review of the data mining applications in medical assessment is provided by Li et al. [2008]. Ekin et al. [2013] provides a discussion of the statistical methods including decision theoretic approaches. Below we refer to payments that result from Fraud, Waste or Abuse, as overpayment.

In general, when the number of claims is small, domain experts audit them one at a time. Comprehensive auditing is beneficial for cases in which overpayments can easily be identified due to irrefutable evidence. For instance, claims that are submitted for dead beneficiaries and simultaneous services to be rendered for the same patient are some examples. There is no need for advanced statistical models for identification of the overpayments for these cases. Descriptive statistical analysis and comprehensive auditing are enough to detect patterns and reveal the potential cases of overpayment.

When the number of claims is larger than what is practically feasible to investigate comprehensively and/or overpayment becomes harder to detect, sampling and estimation methods become necessary. In doing so, there are two important aspects. Firstly, it is important to choose the claims that provide a fair recovery amount without bias. Secondly, the sampling process should provide the claims that help the investigator to reveal the overpayment patterns in the population and inhibit the culture of sustained overpayments. Various techniques are recommended to draw representative samples from the population as efficiently as possible. Designs can be constructed using one or more of sampling techniques such as simple random sampling, systematic sampling, stratified sampling or cluster sampling. In the U.S., use of probability sampling methods for medical investigations has been accepted to be part of the legal framework since 1986. Yancey [2012] provides a comprehensive list about these legal sampling procedures and the parties involved in U.S. governmental medical insurance programs. The population of interest in these procedures is usually the claim payments of which some result with overpayments. According to the current governmental sampling guidelines (CMS [2001]), in most situations the lower limit of a one sided 90 percent confidence interval for the total overpayments should be used as the recovery amount from the provider under investigation. Using the lower bound allows for a reasonable and fair recovery without requiring the tight precision to support the point estimate, sample mean. However, this application of Central Limit Theorem (CLT) is based on the assumption that overpayment population either follows the Normal distribution or that the sample size of overpayments is reasonably large. It is known that medical claims data mostly exhibit skewness and non-normal behavior requiring large sample sizes for the valid application of CLT. It has been shown by Edwards et al. [2003] that methods based on the CLT do not perform well for non-normal overpayment populations where the sample size is small. This may result with some of the investigation decisions being challenged in courts.

The efforts to improve the accuracy of the statistical methods in medical overpayment investigations can be classified in two main areas; representative sampling approaches and overpayment estimation models. As part of the first research stream, over and under sampling approaches are proposed to achieve representative samples (Kubat et al. [1997], Ling and Li [1998]). Under the appropriate conditions, ratio or regression based methods can result in smaller margins of error (CMS [2001]). For instance, Edwards et al. [2003] propose the "Minimum Sum Method", a non-parametric inferential method based on simple random sampling, to obtain a nonrandomized lower bound that is valid for any fixed sample size or any population. Ignatova and Edwards [2008] propose a two stage sampling procedure that can work well for "all or nothing" payment situations. Gilliland and Edwards [2010] construct randomized lower bounds to increase recovery demands. Standard stratified expansion (Buddhakulsomsiri and Parthanadee [2008]) and combined ratio estimators of the total are also proposed to estimate the coverage probabilities of confidence intervals. Our paper employs simple random sampling and focuses on models for overpayment. To the best of our knowledge, the literature on overpayment models is limited. In a relevant work, Mohr [2005] has proposed three different overpayment models and presents simulation results about the conformance of the models with CLT. However, she acknowledges these models are not able to capture the mixed distribution characteristics of medical claims data.

In the five year strategic plan of U.S. Office of Inspector General (OIG [2013]), efforts against overpayments are summarized in three categories; identifying and investigating overpayments, estimating recovery amounts from wrongdoers without bias, and preventing overpayment at the first place. Despite the ground that has been gained in estimating the recovery amounts, there is still a culture of sustained level of overpayments within the Medicare programs. Musal [2010] has brought up the federal budgetary report of "U.S. Office of Management and Budget" that discusses the inability of the existing measures to demonstrate if the health care overpayment has decreased, which is the ultimate mission of the health care initiatives (GPO [2004]). In 2011, an executive of a major consulting firm reckons that because of the sheer volume, overpayment schemes operating at a relatively low dollar value are not investigated at all as they fall below the minimum federal dollar thresholds to prosecute (Gov [2011]). In a press release (CMS [2013]), CMS urges seniors to join the fight against fraud despite the fact that the dollar amounts involved with these transactions is typically smaller. The main objective of these efforts is to ensure that only legitimate providers and suppliers provide services to beneficiaries. Most of the resources are devoted to the investigation of claims with high payments. However, repeated overpayments that result with relatively smaller losses are neglected. This is potentially a result of lack of appropriate overpayment estimation models.

This paper aims to address these concerns and fill the gap in the medical assessment literature by proposing novel overpayment models. In addition to the variants of Mohr's models, we propose a novel inflated mixture based overpayment model that links the known payment population and the information gathered from a sample of investigations.

Our models will help the investigators to understand the overpayment characteristics so that the investigation resources can be allocated to prevent repeated offenses as well as estimating the recovery amounts. Particularly, proposed overpayment models can be beneficial in two ways. Firstly, the estimation of the recovery amount without bias is crucial in standard medical sampling efforts. Proposed models are shown to be robust for particular scenarios via evidence gathered from average coverage probabilities. Secondly, as a unique contribution we estimate the actual overpayment distribution that can be used for identification of the overpayment patterns. Following section introduces the proposed overpayment models. The third section presents the simulation design that considers different scenarios of payment populations and overpayment patterns. It also discusses the estimation methods and performance measures for the proposed models. Section 4 provides a discussion of the results and the paper concludes with a discussion.

2 Overpayment models

In general, the population of payments for medical claims are known, whereas the overpayment values can only be observed after an investigation. The number of claims in a real world scenario can easily reach to a point that is not feasible to investigate each claim. Therefore, we propose models that link payment population and a sample of investigations to estimate the unknown overpayment population. These are referred to as superpopulation models which provide a framework for inference in sampling (Isaki and Fuller [1982]).

Let $\boldsymbol{X} = \{X_1, ..., X_N\}$ be the vector of known payment amounts for a population of N claims and let $\boldsymbol{Y} = \{Y_1, ..., Y_N\}$ represent the unknown overpayment population. As underpayments are very rare in automated claims processing, we assume that an overpayment can take only nonnegative values and cannot exceed its respective payment value. Therefore, the assumed nature of overpayments implies a truncation.

In the medical claims context, two main variables of interest are the percentage of overpayment and the probability of overpayment. Mohr [2005] has estimated the percentage of overpayment in her first two models. She assumes that overpayment follows a Normal distribution with a standard deviation proportional to the functions of the payment value. She provides evidence for these models to work well in the cases of systematic overbilling where the percentage of overpayment is constant. Her third model is constructed to capture the "all or nothing" billing pattern in which the probability of overpayment is estimated. She presents the application of these models for different sampling plans, different estimation techniques and different values of the percentage of overpayment and the probability of overpayment. However, she cautions that these models are not able to capture mixed distribution characteristics. Most of the time, an investigator needs to simultaneously consider fully legitimate, fully and partially overpaid claims to understand the overpayment phenomena. This multimodality of overpayment values and values at extremes is our main motivation to propose an inflated mixture model.

Zero-inflated models have been extensively used in modeling the number of counts with Poisson distribution (Heilbron [2007]). Other applications of inflated Poisson models can be found in Lambert [1992], Gosh et al. [2006] and Neelon et al. [2010]. Ghosh et al. [2006] provide Bayesian inferential results for zero-inflated regression models. We utilize a Beta mixture component to model the percentage of overpayment when it is different from 0 or 1. Beta distribution is a member of exponential family and it is flexible for modeling limited range data, since its density can take different shapes such as left or right skewed. Ospina and Ferrari [2012] provides a discussion about the general class of zero or one inflated Beta regression models.

In the following presentation of the models, we refer to them as Model 1 (M1), Model 2 (M2) and Model 3 (M3). The percentage of overpayment is denoted as ρ whereas the probability of zero overpayment and full overpayment are denoted as π_1 and π_2, respectively. Model 1 (M1) has three main components; claims with no overpayments, claims with overpayment values that are equal to payment values, and claims with partial overpayments. The estimation of ρ, π_1 and π_2 are considered. Model 1 (M1) assumes that the overpayment percentage of the i^{th} claim, ρ_i, is equal to 0 with probability π_1, and is equal to 1 with probability π_2. This overpayment percentage, ρ_i, is observed via the mixture of K Beta distributions with probability $1 - \pi_1 - \pi_2$. Within the mixture with K components; z_k^i, is equal to 1 if i^{th} claim belongs to k^{th} component and its respective k^{th} Beta distribution with parameters (α_k, β_k). Each claim is assumed to belong to only one group at a time. The vector, $\boldsymbol{z}^i = \{z_1^i, \ldots, z_k^i, \ldots, z_K^i\}$, follows a categorical distribution with probabilities $\{\eta_1, \ldots, \eta_k, \ldots, \eta_K\}$. In effect, η_k is the conditional probability that ρ_i is realized from the

k^{th} Beta distribution given that ρ_i is not equal to 0 or 1. $\boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ are the respective matrices with appropriate dimensions. The mathematical notation for Model 1 is provided as;

$$\begin{aligned}
Y_i &= \rho_i X_i \\
P(\rho_i = 0) &= \pi_1 \\
P(\rho_i = 1) &= \pi_2 \\
f(\rho_i | \boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \prod_{k=1}^{k=K} Beta(\alpha_k, \beta_k)^{z_k^i}, \\
z_k^i \in \{0, 1\} \text{ and } \sum_{k=1}^{K} z_k^i &= 1 \text{ and, } \boldsymbol{z}^i \sim Cat(\eta_1, \ldots, \eta_K)
\end{aligned} \tag{2.1}$$

We can compute $\pi_{k+2} = \eta_k * (1 - \pi_1 - \pi_2)$ as the unconditional probability that ρ_i is realized from the k^{th} Beta distribution. The complete probability vector of overpayment patterns, $\boldsymbol{\pi}$ with size $K + 2$ can be written as

$$\boldsymbol{\pi} = \{\pi_1, \pi_2, \pi_3, \ldots, \pi_{k+2}, \ldots, \pi_{K+2}\}$$

In the following section, we explicitly illustrate the models with mixture sizes of $K = 1$ and $K = 2$ and name them as Model 1.A and Model 1.B respectively. It is straightforward to extend this model (M1) and consider different values of mixture sizes, K.

Second and third models below are adapted from the models of Mohr [2005]. Our main difference lies within the estimation technique used for the overpayment percentages. We utilize an investigation sample to estimate it instead of using a predetermined value for simulations.

Model 2, shown below as (2.2), has overpayment following truncated Normal distribution with the error term ϵ_i.

$$\begin{aligned} Y_i &= \rho X_i + \epsilon_i \\ \epsilon_i &\sim TrN(0, \sigma^2 X_i)\mathbb{I}(0 \leqslant Y_i \leqslant X_i) \end{aligned} \tag{2.2}$$

Mohr [2005] has shown that a similar Normal distribution based model can be useful in the cases of systematic fraudulent billing. Our third model, presented below as (2.3), aims to capture the "all or nothing" billing pattern. This can be written as a special case of the proposed M1, when the mixture structure is removed.

$$\begin{aligned} Y_i &= \rho_i X_i \\ P(\rho_i = 0) &= \pi_1 \\ P(\rho_i = 1) &= \pi_2 = 1 - \pi_1 \end{aligned} \tag{2.3}$$

In Model 3, overpayment can be either equal to zero or the respective payment value. This model may be useful for cases either a payment is allowed or denied depending on the beneficiary‘s eligibility.

3 Simulation Design and Estimation Methods

This section presents the proposed simulation design and estimation methods associated with the models explained in Section 2. Different scenarios for payment and overpayment data are considered to display the robustness of the models.

For the payment population, we have evaluated the cases of data following normal and mixed distributions. The publicly available payment data PUF [2010] from Medicare Part B claims is utilized to consider real world medical claims data that exhibit mixed distributional characteristics such as multimodality and values at extremes.

In medical claims context, overpayment population is not available before an investigation, therefore overpayment data are simulated using different patterns. We construct scenarios for the overpayment patterns exhibiting uniform, systematic, mostly legitimate, mostly legitimate or fraudulent (all or nothing) characteristics.

A total of 8 overpayment scenarios are simulated by using 4 overpayment patterns for each of the payment populations. A number of replications of investigation samples are drawn from payment and overpayment data. Using the estimated model parameters, overpayment values for each super-population model are computed.

The performance of each model is assessed by comparing its estimated overpayment values with the overpayment data. **The statistical bias between the actual overpayments and the mean overpayments are reported for each model. The distributional fit between the estimated and actual overpayments is assessed using Kolmogorov-Smirnov (K-S) test. The K-S statistic is based on quantifying a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution (Massey Jr [1951]).** We compute the average recovery amounts of each model when the lower point estimate of 90 percent confidence intervals are used. The average coverage probabilities are also computed to reveal the success of confidence intervals and the application of CLT. Following sub-sections discuss the proposed simulation design and estimation methods in detail.

3.1 Simulation of payment and overpayment data

The first payment population is constructed by retrieving 500 payment values from publicly available payment data PUF [2010] from Medicare Part B claims. It provides mixed distributional characteristics that are common in actual medical claims data, with a mean value of 1251.82 and a standard deviation of 1416.21. Second payment population is drawn from Truncated Normal distribution with mean 300 and standard deviation 25. This provides a symmetric population with 500 instances similar to the one assumed by Mohr [2005]. Figure 1 provides the density plots of these payment populations.

Overpayment patterns are constructed using a pre-determined set of overpayment percentage values and the probability vector, $\boldsymbol{\theta}_j$ for each (j^{th}) scenario. The overpayment percentage takes one of the four values $0, 0.25, 0.75, 1$ with respect to the probability vector, $\boldsymbol{\theta}_j$ for each (j^{th}) scenario. The numerical values of the probability vectors $\boldsymbol{\theta}_j$ are listed in

Table 1.

Overpayment Scenario, $\boldsymbol{\theta}_j$	$\rho = 0$	$\rho = 0.25$	$\rho = 0.75$	$\rho = 1$
$\boldsymbol{\theta}_1$	0.25	0.25	0.25	0.25
$\boldsymbol{\theta}_2$	0.05	0.85	0.05	0.05
$\boldsymbol{\theta}_3$	0.85	0.05	0.05	0.05
$\boldsymbol{\theta}_4$	0.60	0.05	0.05	0.30

Table 1: Probability vectors for simulation of overpayment data with different scenarios

The first overpayment pattern that uses $\boldsymbol{\theta}_1$ provides an example of a structure where all 4 different overpayment percentages can occur uniformly. Second pattern assumes there is high occurrence of claims that has an overpayment percentage of 0.25. Third overpayment pattern has a structure in which most payments do not result with any overpayments, which we also refer to as the zero-inflated structure. With $\boldsymbol{\theta}_4$, we represent the "all or nothing" case where a payment has no overpayment or consists of a full overpayment with respective probabilities of 60% and 30%.

3.2 Estimation of parameters

An investigation sample with size, $n = 60$, is drawn from the payment populations, and the respective sample payment and overpayment values, $\boldsymbol{x} = \{x_1, \ldots, x_i, \ldots, x_n\}$ and $\boldsymbol{y} = \{y_1, \ldots, y_i, \ldots, y_n\}$ are collected.

For the first model, we estimate π_1 and π_2 as $\hat{\pi_1}$ and $\hat{\pi_2}$ using the probability of zero and full overpayments in the sample. We obtain the observed overpayment percentage for i^{th} claim, $\hat{\rho}_i$ via $\frac{y_i}{x_i}$. We proceed to sort the sample and the computed values of $\hat{\rho}_i$ with respect to an ascending order of estimated overpayment percentages. We filter out the cases in which $\hat{\rho}_i$ is either 0 or 1 and proceed to divide the remaining ordered sample into K equal sized mixture components. Given that we have no prior information on which group $\hat{\rho}_i$ belongs to, this method conforms with the maximization of entropy principle (Soofi [1994]) which allows us to handle the potential issue of identifiability. For each mixture component, the mean and standard deviation of the overpayment percentages are computed. The parameters of the k^{th} Beta distribution, α_k and β_k are estimated for the particular sample mean and standard deviation values of each mixture component, k. The auditor can decide on the number of the mixture components, K, using his/her expertise and prior knowledge.

In application for Model 2, ρ is estimated by the average overpayment percentage, $\frac{\sum_{i=1}^{i=n} y_i}{\sum_{i=1}^{i=n} x_i}$ of the sample. For Model 3, we similarly estimate π_1, π_2 using the relative frequency of legitimate and non legitimate payments in the observed sample.

Once the parameters are estimated, the performance of super-population models are compared in estimating the overpayment phenomena as discussed in the next sub-section.

3.3 Measures for model comparison

Firstly, we report the statistical bias and the average recovery amounts for each model with respect to the current *Centers for Medicare and Medicaid Services* (CMS) guidelines. In practice, these guidelines are interpreted so that the lower point estimate of the 90% confidence interval estimate of the overpayments is used in determining the recovery (recoupment) amount. For $10,000$ replications of investigation samples, the average coverage probabilities of the proposed recovery amount are computed. Having a simulation size as large as $10,000$ results with negligible estimation error. This resembles the simulation reports of Mohr [2005], however she presents the average confidence intervals and average coverage probabilities only for fixed values of ρ. We estimate the parameter values using the investigation sample.

Secondly, we evaluate the success of models to detect the characteristics of the overpayment pattern. We compare the overpayment distributions that are estimated using the proposed models, with the overpayment data. This contrasts the current approaches that only consider the mean and standard error. We present density plots and provide p values of Kolmogorov-Smirnov (K-S) tests to assess the distributional fit.

4 Simulation Results

This section presents the simulation results of the proposed models for different combinations of payment data and overpayment patterns. We compare models regarding their performance in estimating the average recovery amount and detecting the overpayment pattern.

4.1 Overpayment scenarios for the mixed payment population

Medical claims data may have mixed distributional characteristics such as the values at extremes and multi-modality. We consider four overpayment scenarios for such payment population. Table 2 provides a summary of the average bias, average coverage probabilities (ACP), average recovery amounts and p-values of K-S tests for each of the proposed models and overpayment scenario.

Overpayment Scenario	Model	Avg. Bias	ACP	Avg. Recovery	K-S p values
1	Model 1.A	38.08	0.9307	490.22	0.39
1	Model 1.B	72.19	0.9337	580.20	0.07
1	Model 2	44.84	0.9269	598.02	0
1	Model 3	284.47	0.9291	755.41	0
2	Model 1.A	15.02	0.9360	284.13	0.03
2	Model 1.B	335.97	0.9321	529.58	0
2	Model 2	39.29	0.9291	269.51	0
2	Model 3	814.15	0.9257	955.80	0
3	Model 1.A	20.93	0.9702	41.42	0.75
3	Model 1.B	19.70	0.9680	44.36	0.86
3	Model 2	22.69	0.9349	68.71	0
3	Model 3	54.06	0.9663	62.05	0.83
4	Model 1.A	37.92	0.9430	334.27	0.32
4	Model 1.B	52.35	0.9381	288.65	0.10
4	Model 2	58.69	0.9274	439.41	0
4	Model 3	78.34	0.9106	375.26	0.11

Table 2: Summary measures for mixed payment population

The first overpayment scenario exhibits 4 major structures that happen uniformly. In terms of average bias, Model 1.A provides the best performance. The K-S test results with p values of 0.39 and 0.07 for M1.A and M1.B, respectively. This suggests we fail to reject the null hypothesis that the estimated overpayment distribution is from the overpayment data for a significance level of 5%. On the other hand, p values of 0 for M2 and M3 provides evidence for rejection of the same null hypothesis. Figure 2 provides further evidence that mixture based models represent the overall overpayment behavior better and they are superior in terms of distributional fit. All the average coverage probabilities being at least 92.69% can be shown as evidence for overcoverage. For Model 3, the high average recovery amount should be treated with caution because of the larger average bias and irrelevance of an all or nothing pattern with this scenario. **The density plots of recovery amounts (see Figure 3) provides evidence that recovery amounts have a bell-shaped curve for each model. It can be recognized that the recovery amount of Model 3 has the highest variance among all four models.**

The second scenario reflects a systematic overpayment structure. According to the results of Mohr [2005], this is the case which Model 2 may be expected to have a good performance. The lower average bias value for Model 2 provides evidence in that direction. However, it should be noted that Model 1.A results with a lower average bias as well as a larger average recovery amount. Overall, average coverage probabilities are relatively high compared to 90% for all models, but Models 2 and 3 are superior. Although the average recovery amounts computed for Models 1.B and 3 are superior, the high average statistical bias suggests the estimation error is larger for these models. Model 1.A has the highest p value of 0.03 indicating the best distributional fit that provides further evidence for its use when the auditor expects a systematic overpayment structure.

The third overpayment scenario represents the case in which most of the claims are legitimate which is a special case of "all or nothing" pattern. Models 1.A and 1.B represent the distributional characteristics of overpayment significantly well with p values 0.75 and 0.86. As expected, Model 3 also has a high p value, 0.83. High p values for K-S test lead to the conclusion that the distributions of estimated and actual overpayments are not statistically different. However, it should be noted that the mixture based models, Models 1.A and 1.B outperform Model 3 with respect to the bias. Model 2 outperforms the others in terms of the average coverage probabilities and recovery amount while having a comparable average bias, although it does not represent the distribution well.

In the case of the fourth overpayment scenario which is similar to the "all or nothing" pattern, Models 1.A, 1.B and 3 are found to provide a good distributional fit as indicated by the p values of 0.32, 0.10 and 0.11. Models 1.A and 1.B outperform the other models with respect to the average bias albeit with larger average coverage probabilities.

4.2 Overpayment scenarios for the symmetric payment population

For the symmetric payment population, Table 3 provides a summary of the average bias, average coverage probabilities (ACP), average recovery amounts and p-values of K-S tests for each of the proposed models and overpayment scenario.

Overpayment Scenario	Model	Avg. Bias	ACP	Avg. Recovery	K-S p values
1	Model 1.A	8.64	0.9117	138.19	0
1	Model 1.B	25.84	0.9070	157.41	0
1	Model 2	8.29	0.9187	154.12	0
1	Model 3	84.32	0.8990	213.12	0
2	Model 1.A	10.62	0.9200	70.99	0
2	Model 1.B	81.07	0.9181	149.98	0
2	Model 2	10.39	0.9140	75.49	0
2	Model 3	185.10	0.8688	260.31	0
3	Model 1.A	7.89	0.9435	21.34	0.41
3	Model 1.B	9.11	0.9420	22.36	0.40
3	Model 2	8.12	0.9174	33.74	0
3	Model 3	16.50	0.9338	27.24	0.13
4	Model 1.A	18.14	0.9191	102.15	0.002
4	Model 1.B	20.71	0.9266	68.27	0
4	Model 2	20.09	0.9160	124.06	0
4	Model 3	36.81	0.9163	119.35	0

Table 3: Summary measures for symmetric payment population

In the case of the first scenario which has a systematic overpayment pattern, Models 1.A and 2 result with lower values of average bias. In terms of representing overall overpayment behavior; all the models have p-values of 0 for the K-S test. **Visual evidence gathered from Figure 4 indicate that the inflated mixture models are slightly superior. This may signal that the mixture models with larger mixture sizes may provide a better fit. Despite slightly higher average coverage probability, Model 2 outperforms Model 1.A with respect to the average recovery amount.** Model 3 provides the best average recovery amount and the average coverage probability. However, that should be treated with caution because of the relatively larger bias and irrelevance of an all or nothing pattern.

The second scenario having one major overpayment structure leads to the expectation of Model 2 performing the best. The simulation results confirm this since Model 2 provides the best bias, average coverage probability and average recovery amount. Figure 5 shows that Models 1.A, 1.B and 2 have mixed success to represent the overpayment patterns wheras p values of the K-S test for all models are found to be 0. In addition to having a relatively very high bias, Model 3 also results with undercoverage for this pattern.

The third overpayment scenario represents the case in which most of the claims are legitimate. The overall performance of the models are relatively better compared to the performance within other scenarios. Model 1.A results with the best bias, wheras Model 2 provides the best average coverage probability and recovery amount. Models 1.A and 1.B represent the distributional characteristics of overpayment significantly well with reported p values of K-S test as 0.41 and 0.40. However, Model 2 outperforms the others in terms of the average coverage probabilities and the recovery amounts. **Although it does not represent the distribution well enough (Figure 6), it represents the point estimates better than other models.**

The fourth scenario is very similar to the “all or nothing” case with other patterns emerging only 10% of the time. Model 1.A provides the lowest bias whereas the average coverage and recovery amounts of Model 2 are better. However, Figure 7 suggest that Model 2 does worse in terms of distributional fit compared to all other models. K-S test result with p values that are 0 or very close to 0, therefore we cannot suggest none of the models provide a good distributional fit.

5 Conclusion

Sampling approaches prove useful when the number of claims is large and it is impractical to audit all claims. Existing methods are known to be successful for particular scenarios in terms of estimating the recovery amounts while conforming with the sampling guidelines. However, as stated by federal governmental agencies, this has not inhibited the sustained level of overpayments within the Medicare programs. Detecting the overpayment patterns still remains as a problem. The proposed models can be utilized for both of these purposes;

estimating recovery amounts while also identifying the overpayment patterns. The investigators can benefit from these models in particular scenarios by using their expertise and initial investigation results.

We investigate the performance of the models for different overpayment patterns and payment populations. For the simulations with mixed payment population, inflated mixture models are found to provide superior distributional fit with a smaller average bias. The inflated mixture model is flexible and is able to capture the skewness and non-normality of the payment data. Using the current sampling guidelines that require a point estimate, mixture models also result with a higher recovery amount and has a more precise fit to the hypothetical average coverage probability compared to Model 2. The average coverage probabilities indicate over-coverage for particular scenarios. Assessing the distributional fit helps us to understand the benefits of each model to capture the overall overpayment pattern. This may mean that the investigator ends up auditing more claims that have medium or lower payment levels so that the final sample is more balanced and the mixed distribution characteristics are captured. Committing more resources against overpayment at the medium payment levels may prevent repeated overpayments. Overpayment models with better distributional fit also provides statistical support in potentially requesting a higher recovery amount than the point estimate.

The p values that are used to assess the distributional fit are 0 **for most scenarios with the symmetric payment population. This implies that we reject that overpayment sample does not fit the actual overpayment data. Mixture models with larger mixture sizes are expected to perform better with respect to the distributional fit. An ideal mixture size depends on the data characteristics such as the homogeneity of the payment population and overpayment patterns. For instance, if the decision maker has prior expertise suggesting the data set of concern involves various patterns, then he/she may want to consider models with larger mixture sizes depending on the suspected number of patterns. Whereas if it is anticipated that the data set includes one pattern and is fairly homogeneous, a mixture model with one mode or a normality based model can be good enough for estimation. A larger mixture size can provide better fit, but the modeler**

should also be aware of the additional complexity it brings and needs to evaluate the trade-off using a model selection criteria. Future work may include the incorporation of the consideration of the unknown mixtures size, K, within the model selection.

The use of these overpayment models can be extended to stratified samples. In a stratified sampling setting proposed methods are applicable to each stratum. Current stratification approaches include the cumulative root frequency method of Dalenius and Hodges Jr [1959] and the method of Lavallée and Hidiroglou [1988] for skewed populations (See Cochran [2007] for a detailed discussion). Basically, you can divide the population into S groups containing $N_1, N_2, ..., N_S$ units based on the sorted payment values. Then the overpayment values within each group can be estimated using the proposed overpayment models. Another extension would be to include covariates to improve the estimation for the overpayment models. Consideration of a random shock within the mixture models can capture more variability which may result with improved estimates.

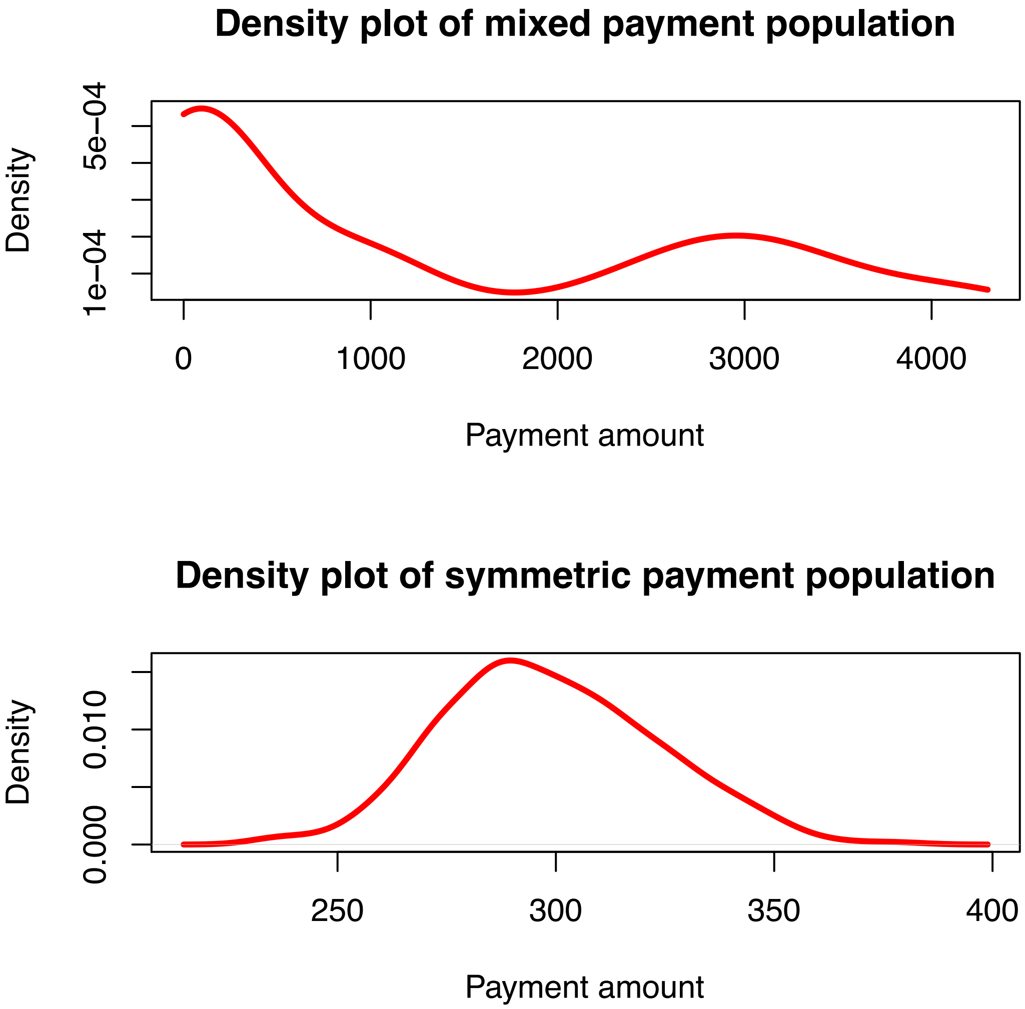

Figure 1: Density plots of payment populations

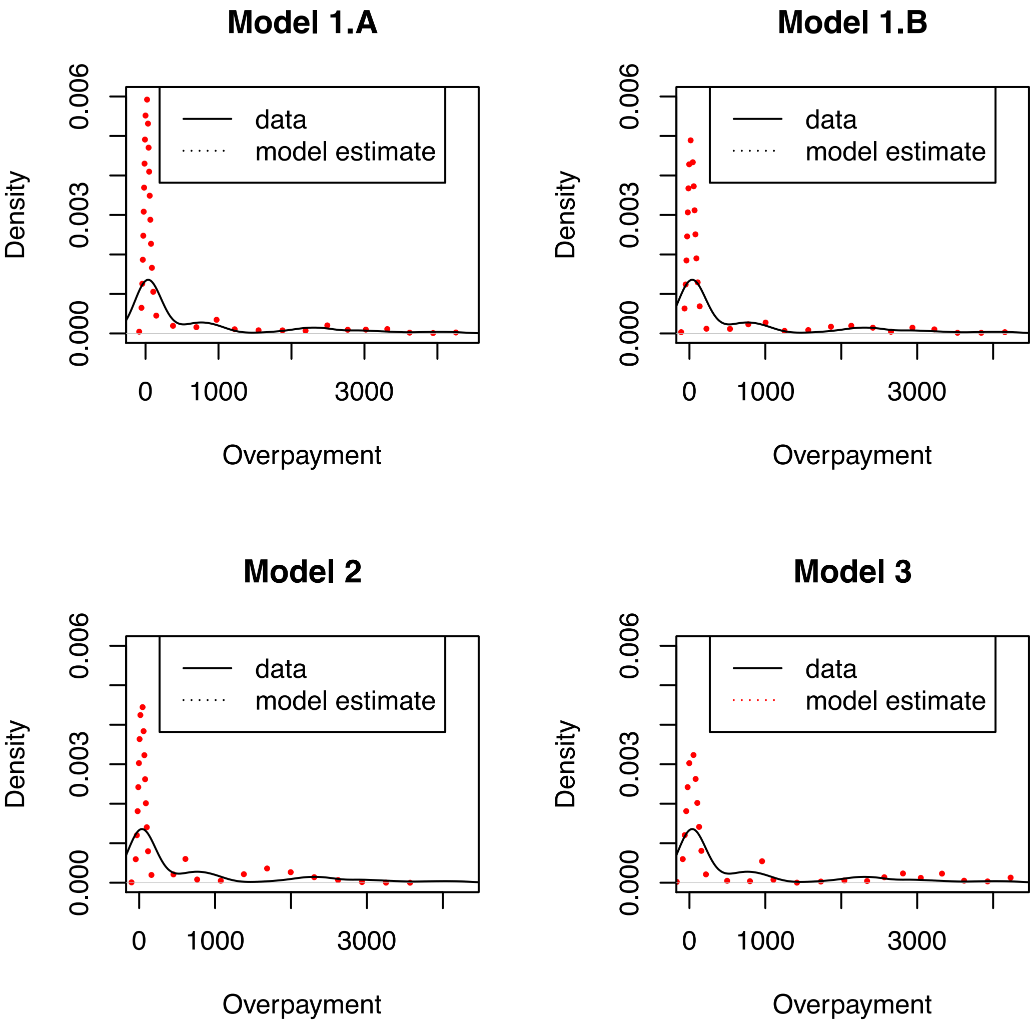

Figure 2: Density plots of actual and estimated overpayments with mixed payment population and first overpayment pattern

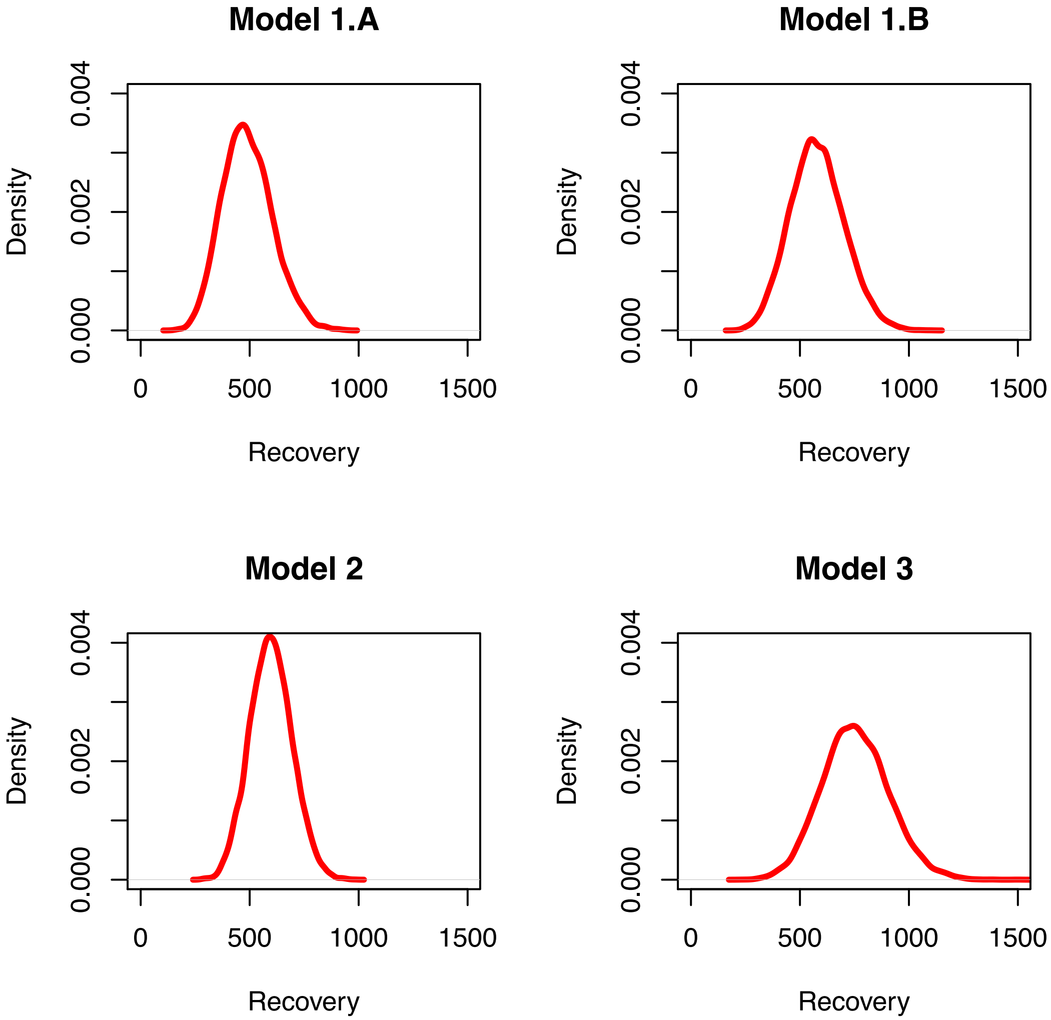

Figure 3: Density plots of the recovery amounts with mixed payment population and first overpayment pattern

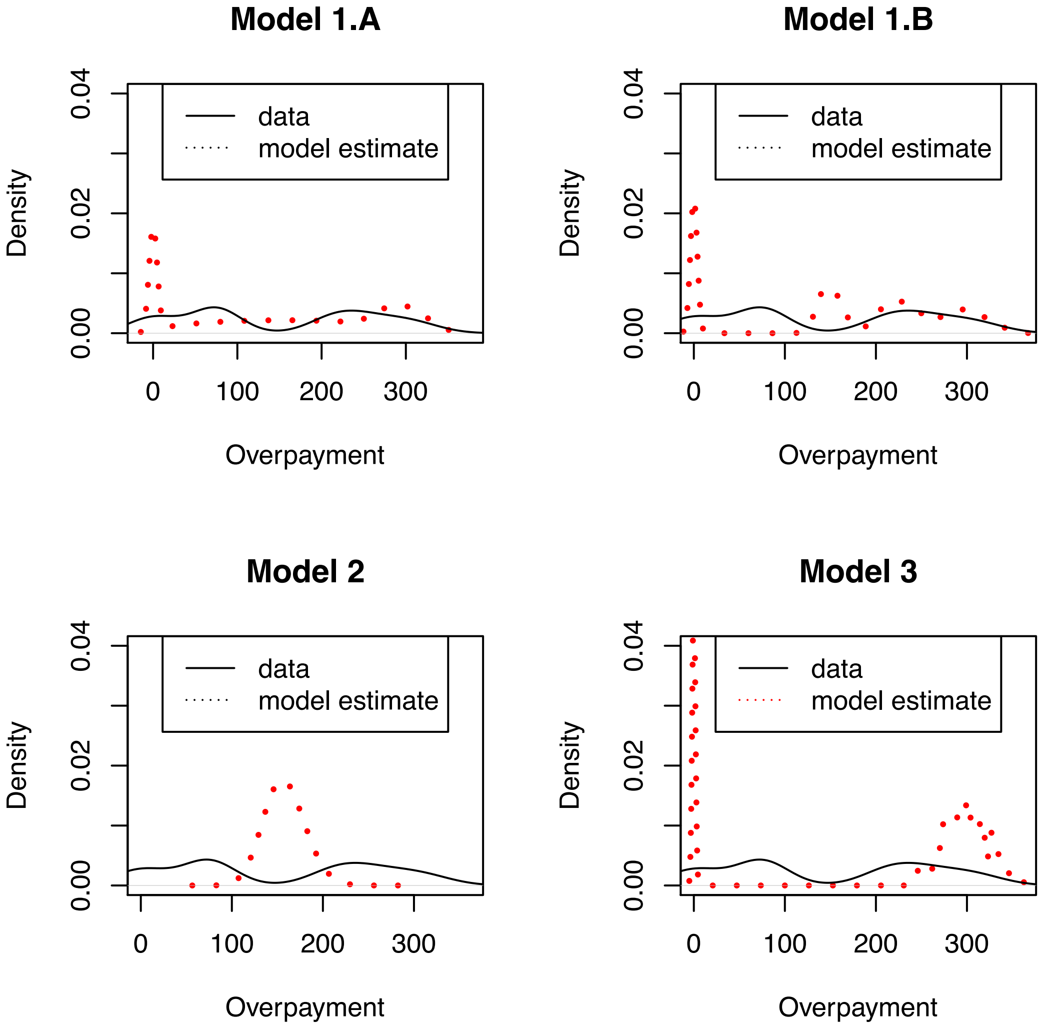

Figure 4: Density plots of actual and estimated overpayments with symmetric payment population and first overpayment pattern

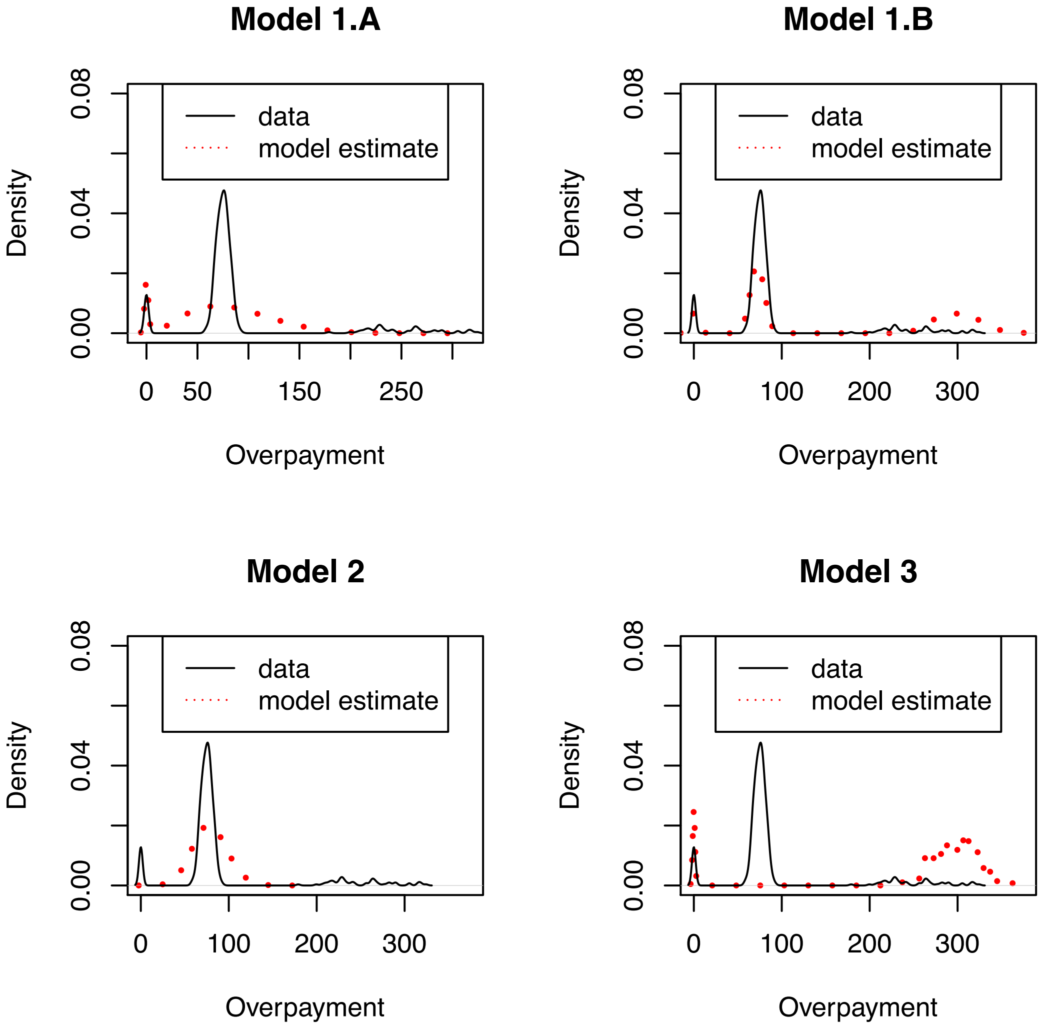

Figure 5: Density plots of actual and estimated overpayments with symmetric payment population and second overpayment pattern

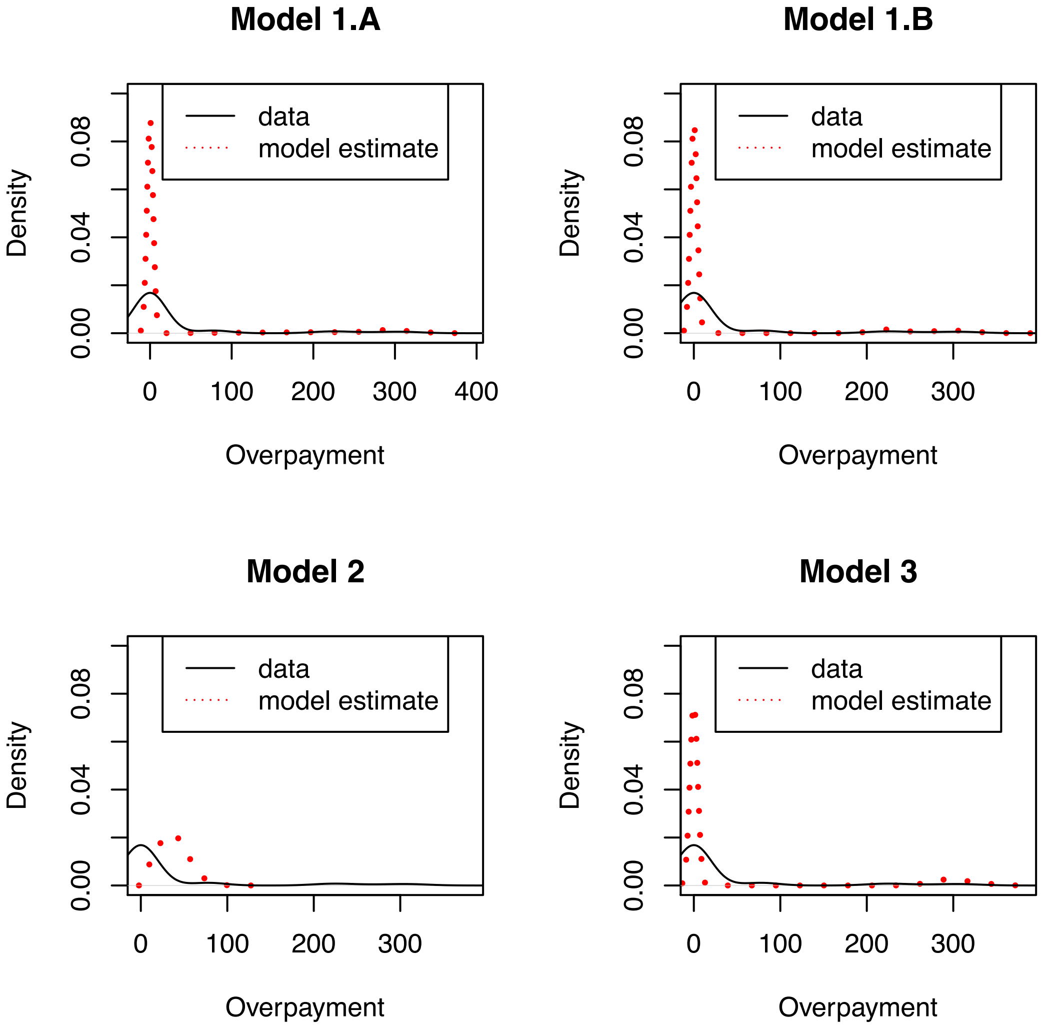

Figure 6: Density plots of actual and estimated overpayments with symmetric payment population and third overpayment pattern

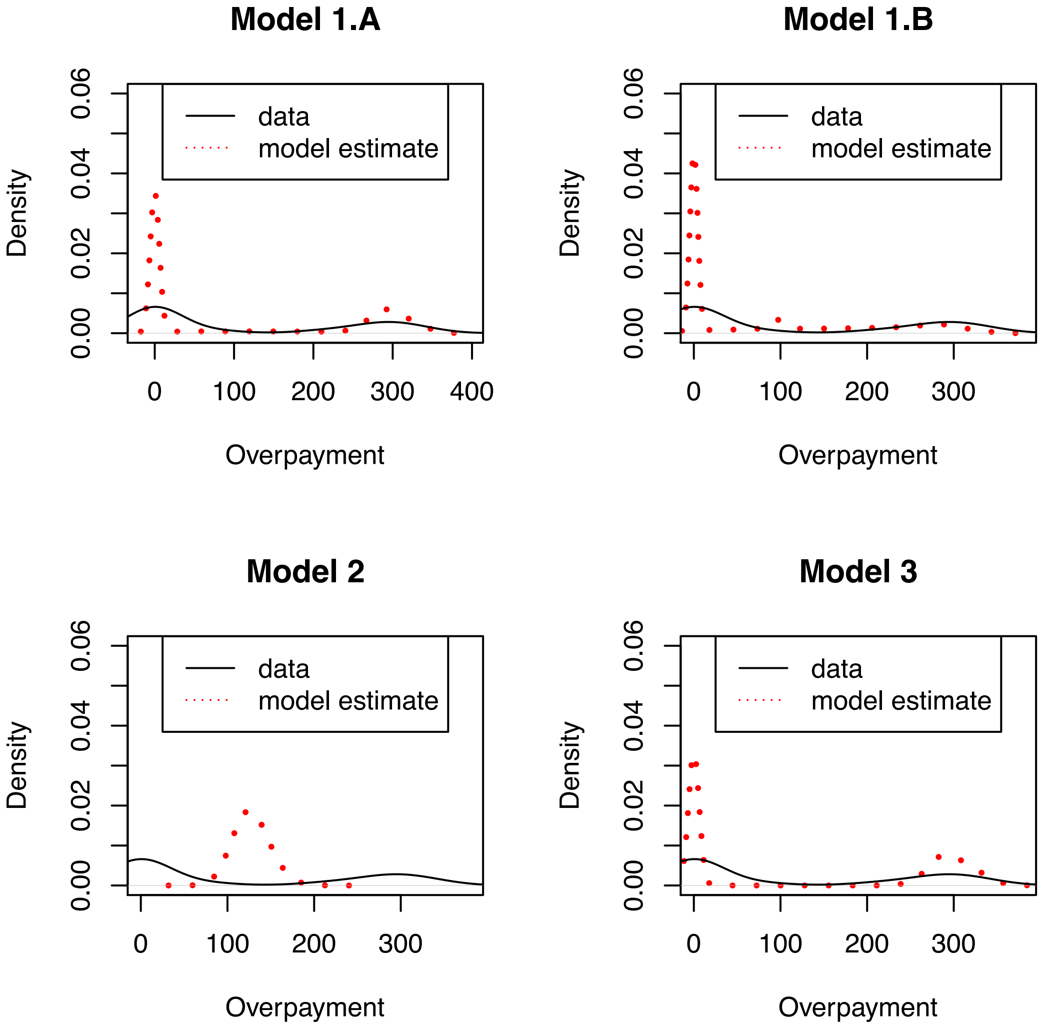

Figure 7: Density plots of actual and estimated overpayments with symmetric payment population and fourth overpayment pattern

References

Program memorandum carriers transmittal b-01-01. http://www.cms.gov/Regulations-and-Guidance/Guidance/Transmittals/downloads/B0101.pdf, 2001. Accessed: 01/10/2015.

Budget of the united states government, fiscal year 2004: Performance and management assessments. http://www.gpo.gov/fdsys/pkg/BUDGET-2004-PMA/pdf/BUDGET-2004-PMA.pdf, 2004. Accessed: 01/10/2015.

Cms financial report 2010. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/CFOReport/Downloads/2010_CMS_Financial_Report.pdf, 2010. Accessed: 01/10/2015.

Basic stand alone (bsa) medicare claims public use files (pufs). http://cms.hhs.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/BSAPUFS/index.html, 2010. Accessed: 01/10/2015.

It‘s all in the details. http://www.govconexec.com/2011/06/01/its-all-in-the-details/, 2011. Accessed: 01/10/2015.

Details for title: Medicare urges seniors to join fight against fraud. http://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-Releases/2013-Press-Releases-Items/2013-06-06.html, 2013. Accessed: 01/10/2015.

Oig strategical plan 2014-2018. https://oig.hhs.gov/reports-and-publications/strategic-plan/files/OIG-Strategic-Plan-2014-2018.pdf, 2013. Accessed: 01/10/2015.

G. Anderson and P.S. Hussey. Comparing health system performance in oecd countries. *Health Affairs*, 20(3):219–232, 2001.

J. Buddhakulsomsiri and P. Parthanadee. Stratified random sampling for estimating billing accuracy in health care systems. *Health Care Management Science*, 11(1):41–54, 2008.

William G Cochran. *Sampling techniques*. John Wiley & Sons, 2007.

Tore Dalenius and Joseph L Hodges Jr. Minimum variance stratification. *Journal of the American Statistical Association*, 54(285):88–101, 1959.

D. Edwards, G. Ward-Besser, J. Lasecki, B. Parker, K. Wieduwilt, F. Wu, and P. Moorhead. The minimum sum method: a distribution-free sampling procedure for medicare fraud investigations. *Health Services and Outcomes Research Methodology*, 4(4):241–263, 2003.

Tahir Ekin, Francesca Ieva, Fabrizio Ruggeri, and Refik Soyer. Statistical issues in medical fraud assessment. Technical report, The George Washington University The Institute of Integrating Statistics in Decision Sciences, 2013.

Sujit K Ghosh, Pabak Mukhopadhyay, and Jye-Chyi JC Lu. Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, 136(4):1360–1375, 2006.

Dennis Gilliland and Don Edwards. Using randomized confidence limits to balance risk an application to medicare fraud investigations. *Statistics and Probability Research Memorandum RM-685, Michigan State University*, 2010.

Sujit K. Gosh, Pabak Mukhopadhyay, and Jye-Chyi Lu. Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, 136(4):1360–1375, April 2006.

David C Heilbron. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36(5):531–547, 2007.

Iliana Ignatova and Don Edwards. Probe samples and the minimum sum method for medicare fraud investigations. *Health Services and Outcomes Research Methodology*, 8(4):209–221, 2008.

Cary T Isaki and Wayne A Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982.

M. Kubat, S. Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Machine Learning-International Workshop then Conference*, pages 179–186. Morgan Kaufmann Publishers, Inc., 1997.

Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

Pierre Lavallée and M Hidiroglou. On the stratification of skewed populations. *Survey Methodology*, 14(1):33–43, 1988.

Jing Li, Kuei-Ying Huang, Jionghua Jin, and Jianjun Shi. A survey on statistical methods for health care fraud detection. *Health Care Management Science*, 11:275–287, 2008.

C.X. Ling and C. Li. Data mining for direct marketing: Problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 73–79, 1998.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

Donna L. Mohr. Confidence limits for estimates of totals from stratified samples, with application to medicare part b overpayment audits. *Journal of Applied Statistics*, 32(7): 757–769, 2005.

R.M. Musal. Two models to investigate medicare fraud within unsupervised databases. *Expert Systems with Applications*, 37(12):8628–8633, 2010.

Brian H. Neelon, A. James O'Malley, and Sharon-Lise T. Normand. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*, 10(4):421–439, 2010.

Raydonal Ospina and Silvia LP Ferrari. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623, 2012.

Hyunjung Shin, Hayoung Park, Junwoo Lee, and Won Chul Jhee. A scoring model to detect abusive billing patterns in health insurance claims. *Expert Systems with Applications*, 39 (8):7441–7450, 2012.

Ehsan S Soofi. Capturing the intangible concept of information. *Journal of the American Statistical Association*, 89(428):1243–1254, 1994.

Will Yancey. Sampling for medicare and other claims. http://www.willyancey.com/sampling-claims.html, 2012. Accessed: 01/10/2015.