Information Theoretic Multi-Stage Sampling Framework for Medical Audits

Muzaffer Musal and Tahir Ekin *

March 25, 2018

Abstract

The sampling resource allocation decisions for medical audits of outpatient procedures are crucial and challenging because of the large payment amounts and heterogeneity of the claims. A number of frameworks are utilized to help auditors to address the trade-offs between efficiency and cost while having valid overpayment amount estimates. As a potential improvement, this paper presents a novel information theoretic multi-stage sampling framework. Particularly, we propose an iterative stratified sampling method that uses Lindley's entropy measure to evaluate the expected amount of information. We use U.S. Medicare Part B claims outpatient payment data and investigate the versatility of the framework for different overpayment scenarios and resource allocation designs. The proposed method results in reasonable coverage and lower estimation errors for proportion of overpaid claims and overpayment recovery amounts. Our sampling method is shown to outperform the current stratification method of practice, Neyman Allocation for many scenarios. The framework also can be used to make probability statements on variables of interest, such as number of overpaid claims.

Keywords: Medical claims; Medical audits; Medicare; Stratified sampling; Entropy; Overpayment estimation; Bayesian methods.

*Both authors are at McCoy College of Business, Texas State University. Dr. Ekin is the corresponding author with the contact information 601 University, McCoy 451, San Marcos Texas 78666, 001 512 245 3197,t_e18@txstate.edu.

1 Introduction

Healthcare spending is a significant item in the governmental budgets, especially in developed countries. In the United States, this reached \$3.2 trillion or \$9,990 per person, which accounted for 17.8 percent of the nation's Gross Domestic Product (CMS (2017)). Particularly, the number of outpatient procedures has increased remarkably tripling over the past 30 years to more than 54 million a year. By the end of 2017, the percentage of outpatient procedures is estimated to be 75 percent of overall procedures (Jeffries (2016)). Outpatient procedures correspond to cases where patients come to the provider for diagnosis, treatment or surgery but are not admitted for an overnight stay. Their advantages may include faster recovery and improved patient safety.

U.S. governmental agencies report that each year three to ten percent of the overall health care spending is lost to fraud, waste and abuse (Shin et al. (2012)). These overpayments impact both the government and tax-payers and have direct cost implications. They can also diminish the ability of medical systems to provide quality care to beneficiaries. In order to determine the legitimacy of the claim submissions and identify overpayments, medical audits are conducted. This requires subject domain experts to manually investigate the data of claims, providers and beneficiaries, which is generally costly and time-consuming. The size of the medical systems make comprehensive auditing unfeasible, therefore sampling methods are necessary. Managing the trade-off between the audit costs and accuracy of the extrapolations and overpayment estimation is one of the main challenges in sampling and subsequent resource allocation decisions. The investigation costs correspond to the time spent by the experts as well as the physical resources. Potential audit of claims that have zero overpayment (false positives) can result in a loss of trust in the government and a lost opportunity cost. This is especially important for outpatient procedures, where the heterogeneity of the procedures is relatively large. Such heterogeneous nature of medical claims data and the existence of a number of fraud patterns can increase the estimation error. Even a small improvement in the sampling resource allocation while satisfying the governmental guidelines is crucial.

Most of the medical audit sampling approaches are performed in a single stage, see Ekin et al. (2018) a comprehensive overview. To the best of our knowledge, the only multi-stage medical audit sampling approach is proposed by Ignatova and Edwards (2008). The population of interest in these sampling procedures are usually the payment amounts to a provider, of some of which consist of 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 can be used as the recovery amount from the provider under investigation. Using the lower bound aims to protect the government from recovering an amount greater than the true value of overpayments. The simple expansion, ratio and regression-based estimators are among the widely used estimation methods in audits; see Guthrie et al. (1989) for an overview. Medical claims data is well known to exhibit skewness and non-normal behavior, thus, requiring large sample sizes for accurate estimation. Alternative approaches include but are not limited to the minimum-sum method of Edwards et al. (2003), the zero-one inflated mixture model of Ekin et al. (2015) and the Bayesian mixture model of Musal and Ekin (2017).

This paper proposes an information theoretic multi-stage stratified sampling framework for audits of U.S. Medicare Part B outpatient claims. In doing so, the proposed approach uses Lindley's entropy measure in a Bayesian framework to quantify the expected information content of sampling from each stratum. Additionally, the validity of the overpayment recovery estimates with respect to the governmental guidelines is investigated. The performance of the proposed method is then compared to the current stratification method of practice, the Neyman Allocation (Neyman (1934)) for a number of overpayment scenarios and allocation designs.

The contributions of this paper are three-fold. First, from the methodological perspective, this is a novel multi-stage sampling framework in that the auditor evaluates the information content in each step and makes sampling resource allocation decisions using the expected information gain. Although it is motivated and illustrated for medical audits, the proposed approach is general in that it can be adapted to other domains that may benefit from

iterative sampling. Second, the proposed method is shown to be a valid and efficient sampling framework for real world multi-modal medical claims payment data and various overpayment scenarios assuming fully fraudulent or fully legitimate (all or nothing) claims. It outperforms the Neyman Allocation method with respect to the estimation precision for the proportion of overpaid claims and the overpayment recovery amount. The proposed method also provides estimates for each stratum. This can help medical auditors allocate sampling resources among strata in addition to retrieving valid overpayment estimates. Third, this framework can help auditors quantify the uncertainty concerning variables of interest via their respective probability distributions. For instance, the probability distribution of number of overpaid claims can help understand overpayment patterns. This can help accomplish the ultimate mission of the health care initiatives, decreasing overall overpayments.

The paper is organized as follows. The following section introduces the use of entropy as an information measure. Section 3 outlines the proposed method in detail. Section 4 describes the medical outpatient claims data with an emphasis on stratification design and the generation of overpayment scenarios. Section 5 illustrates the use of the method with an analysis, and the paper concludes in Section 6.

2 Lindley's Entropy as an Information Measure

Entropy has been used to quantify the available information in a number of different disciplines. Outside of thermodynamics, the first major application of entropy was developed by Shannon (1948) in the domain of information theory. In the context of communication, distributions with larger deviations indicate relatively more information, motivating Shannon's entropy. Lindley (1956) has pointed out that the opposite can be true in statistics. This interpretation of informational value is referred to as Lindley's entropy. Soofi (1994) describes entropy as the uncertainty summary about a probability distribution and defines it as the expected value of the log probability distribution. He presents an overview of different interpretations of entropy within the statistics literature. Ebrahimi et al. (2010) list duration analysis, order statistics, data disclosure and evaluation of predictor variables as

four potential areas that entropy is used as an information measure. Information measures have also been utilized extensively within Bayesian experimental design problems, see the review of Chaloner and Verdinelli (1995). Particularly, entropy has been used in algorithms such as maximum entropy sampling (Shewry and Wynn (1987)) which aims to choose a most informative subset of random variables. Other applications include quantification of information amount in the context of multicriteria decision problems (Musal and Soyer (2014)). In this paper, we will use entropy as an information measure within a stratified sampling framework.

First, we provide the mathematical background for the use of entropy as an information measure. Let's define S as the random variable of interest, s as the sample (data) and θ as the parameter of interest. The posterior distribution of θ, $p(\theta|s)$, is obtained as proportional to the product of the likelihood $L(s|\theta)$ and prior density $p(\theta)$ using the Bayes rule:

$$p(\theta|s) = \frac{L(s|\theta)p(\theta)}{\int_\theta L(s|\theta)p(\theta)} \tag{2.1}$$

After obtaining $p(\theta|s)$, the posterior predictive distribution of S is written as

$$p(S|s) = \int_\theta p(S|\theta, s)p(\theta|s)d\theta. \tag{2.2}$$

In cases where $p(S|s)$ cannot be obtained analytically, it can be approximated using the Monte Carlo average as in

$$p(S|s) \approx \frac{1}{G}\sum_{g=1}^{g=G} p(S|\theta^{(g)}), \tag{2.3}$$

by using G samples of $\theta^{(g)}$ that are retrieved from $p(\theta|s)$ assuming the conditional independence of S and s given θ.

Using entropy, the information content of the random variable S at a given iteration t can be quantified as:

$$g_t = \int_S p(S)log(p(S))dS \tag{2.4}$$

where $p(S)$ is the prior predictive distribution of S. After the collection of the sample s, the

uncertainty of S is updated via $p(S|s)$, and the information content of S at iteration $t+1$ becomes

$$g_{t+1} = \int_S p(S|s) log(p(S|s)) dS, \tag{2.5}$$

Lindley (1956) quantifies the change in information content of S after observing s via the entropy function $\Delta g(s) = g_{t+1} - g_t$. This gain can be negative or positive (Soofi (2000)). Potential negativity of $\Delta g(s)$ would imply increased uncertainty of S after observation of sample s.

Furthermore, the expected informational value of a prospective observation can be evaluated. This is referred to as Lindley's expected information gain and is equivalent to the expected *Kullback Leibler* divergence of prior and posterior distributions. This measure is always non-negative (Musal and Soyer (2014)). It should be noted that, unlike variance, entropy has an additive property which makes it beneficial for iterative algorithms.

Once we observe sample s at iteration $t+1$, the expected information gain from a prospective single observation s_f at iteration $t+2$ is computed as:

$$E[\Delta g(s)] = E_{s_f}[g_{t+2}] - g_{t+1} \tag{2.6}$$

This can be written as

$$\begin{aligned} E[\Delta g(s)_{t+1,t+2}] = &\int_{s_f} \int_{s_{ff}} p(s_f|s) p(s_{ff}|s, s_f) log(p(s_{ff}|s, s_f)) ds_f ds_{ff} \\ &- \int_{s_{ff}} p(s_{ff}|s) log(p(s_{ff}|s)) ds_{ff}, \end{aligned} \tag{2.7}$$

where s_{ff} is an additional prospective sample. In our setting, the s_{ff} term will be useful in comparing the expected information gain among strata. For each stratum, we assume s_f as the observed sample and evaluate how it would affect the expected information gain via the probability distribution of every possible value evaluated at s_{ff}. The term $p(s_{ff}|s, s_f)$ is the uncertainty distribution of S, evaluated at all possible values, given s and s_f. In the following, we will benefit from this idea to evaluate the stratum with the highest expected

information value and therefore will be sampled from.

3 Methodology for a Stratified Sampling Framework

This section outlines the steps for the proposed method of applying Lindley's entropy within a multi-stage sampling framework. In addition we also provide a subsection on the method of comparison for the proposed method's results versus the Neyman Allocation's results.

In medical audits, the variables of interest are the proportion of overpaid claims, ρ, and number of overpaid claims in the population, K, as well as the total overpayment and recovery amounts. Next, using entropy as an information measure and retrieving the probability distributions of ρ and K, we propose an iterative stratified sampling framework for all or nothing claims. The proposed framework is based on sampling from the stratum with the highest uncertainty of the number of overpaid claims and provides estimates of overpayment recovery amount.

First, we introduce the following notation. The total number of claims in the population is $N = \sum_{h=1}^{L} N_h$ where N_h is the number of claims in stratum h. The unknown total number of overpayments in the population is denoted as $K = \sum_{h=1}^{L} K_h$ where K_h is the number of overpaid claims in stratum h. The payment and overpayment amounts of claims in stratum h are represented by the vectors $\boldsymbol{X}_h = \{X_{(h,1)}, ..., X_{(h,N_h)}\}$ and $\boldsymbol{Y}_h = \{Y_{(h,1)}, ..., Y_{(h,N_h)}\}$, respectively. We assume that $\boldsymbol{X}_h$ is known but $\boldsymbol{Y}_h$ is only known after investigation. $\boldsymbol{\rho}$ denotes the proportion of overpaid claims, a vector of size L, that consists of ρ_h, the proportion of overpaid claims in each stratum h.

The symbol k denotes the number of overpaid claims in a sample of size n. Once the samples are allocated to strata, the sample payment and overpayment amounts of claims in stratum h are represented by the vectors $\boldsymbol{x}_h = \{x_{(h,1)}, ..., x_{(h,n_h)}\}$ and $\boldsymbol{y}_h = \{y_{(h,1)}, ..., y_{(h,n_h)}\}$, respectively. The number of overpaid claims in stratum h is denoted as k_h in a sample of n_h. The total number to be sampled at t^{th} iteration is $n(t)$ whereas the number of claims to be sampled from stratum h at iteration t is denoted as $n_{h,t}$, and the number of overpaid claims in stratum h at iteration t is $k_{h,t}$.

The steps of the proposed framework are listed as follows:

1. Data pre-processing
   (a) Determine the number of strata, L, and stratum boundaries
   (b) Allocate payments to each stratum with respect to the stratum boundaries
2. Initial resource allocation at iteration $t = 0$
   (a) Perform an initial sample allocation (determine $n_{h,0}$, sample size of each stratum at 0^{th} iteration, $h = 1, ..., L$) using Neyman Allocation (N.A.)
   (b) Increase the iteration value by 1 and set it as $t = 1$
3. Additional iterative allocation: Repeat until $t = T$ where T is the last iteration
   (a) For stratum h, at t^{th} iteration; obtain $p(\rho_{h,t}|k_{h,t}, n_{h,t})$, the posterior distribution of the proportion of overpaid claims conditional on the draw of an additional sample, $n_{h,t}$ and outcome if that is legitimate or overpaid, $k_{h,t}$
   (b) For stratum h, at t^{th} iteration; obtain the distribution $p(K_{h,t}|k_{h,t})$
   (c) For stratum h, at t^{th} iteration; obtain $p(k_{h,t}|n_{h,t})$ and compute $E[\Delta_h(k_f)]$, the expected information gain for an additional sample k_f.
   (d) At iteration t, determine $h^*(t)$, the stratum with the highest expected information gain and sample from $h^*(t)$
4. At iteration T, estimate the proportion of overpaid claims, ρ and the overpayment recovery amounts for the all strata as well as the overall population.

3.1 Step 1: Data pre-processing

The first step involves data-preprocessing. This includes the determination of number of strata, L and the strata boundaries. We choose the number of strata as $L = 5$, since CMS is

documented to approve the use of five or six strata (Park and Perling (2016)). It should be noted that designs that have a different number of strata do not make the analysis invalid.

There are a number of strategies that can be used to determine the stratum boundaries. These include the cumulative square root rule (Dalenius and Hodges Jr (1959)) and the stratification approach of Lavallée and Hidiroglou (1988) for skewed populations. Horgan (2006) provides a survey of such methods and proposes a geometric progression based method to determine the stratum boundaries. This method is shown to be free of convergence problems and arbitrariness of initial values. Hidiroglou and Kozak (2017) compare the optimization based and approximate methods, and recommend the use of optimization based methods such as Kozak (2004) since their complexity issues are addressed by recent software developments. Any of these strategies can be used within the proposed framework.

Lastly, with respect to chosen stratum boundaries, the claims and the respective payment values are allocated to each stratum.

3.2 Step 2: Initial resource allocation

The initial resource allocation is crucial and depends on the context and objective of the analysis. The step of determining the sample sizes of each stratum strongly influences the precision and cost of a stratified sampling design. Most widely used designs are proportional and disproportional allocations. Proportional allocation corresponds to cases in which the size of the sample drawn from each stratum is proportional to the relative size of the stratum in the population. However, this may not be efficient when the group of interest does not have enough claims in a sub-group of the population. Therefore a number of disproportional stratified sampling methods are proposed depending on the goal of the auditor. For instance, if the aim is to focus on a given stratum, the auditor may want to ensure that the samples drawn from that stratum are large enough. If the objective is to conduct between-strata analysis, then an equal number of samples can be chosen for each stratum. As an alternative, the optimum allocation method allows the auditor to consider both the precision (variance) of the estimates and the sampling cost. The objective is to sample more heavily from a

stratum when the cost to sample an element from the stratum is low, the variance within the stratum is large, and the population size of the stratum is large. When the costs are not available or when they are equal, optimum allocation can be modified so that it only considers the variance within each stratum. This is referred to as the Neyman Allocation, and is shown to be an efficient stratified sampling method (Mathew et al. (2014)). While we use Neyman Allocation in this manuscript for initial resource allocation in this manuscript, other design can also be embedded to the proposed framework.

For the Neyman allocation method, a predetermined number, $n(0)$, of units are allocated to strata in the initial iteration (t=0) based on the standard deviation, σ_{X_h} and mean, $\hat{X}_h$ payment of each stratum, $h = 1, ..., L$. The sample size for stratum h can be written as:

$$n_{h,0} = \frac{n(0)N_h\sigma_h}{\sum_{h=1}^{h=L} N_h\sigma_h} \tag{3.1}$$

where

$$\sigma^2_{X_h} = \frac{\sum_{i=1}^{i=N_h} (X_{h,i} - \hat{X}_h)^2}{N_h - 1}$$

It should be noted that payment is not the actual variable of interest. However, the actual measurements of overpayments are not known before an audit, therefore we assume overpayments are strongly linked to the payment values. This can be a naive assumption especially in the cases with partial overpayments.

3.3 Step 3: Additional iterative allocation

Next, we obtain the posterior distribution of $\rho_{h,t}$, the proportion of overpaid claims for stratum h at iteration t. We suppress h and t for clarity and only reintroduce them when necessary. The probability distribution of k is Binomial with parameters (n, ρ) and the prior distribution of ρ is defined via Beta distribution with the hyper parameter vector $\boldsymbol{\omega} = (\alpha, \beta)$. Due to conjugacy, this leads to the posterior distribution that is needed in Step (3a) as:

$$p(\rho|k,n) \sim Beta(\alpha+k, \beta+(n-k)). \tag{3.2}$$

The posterior distribution of the number of overpaid claims in the population, K, in a particular stratum can be evaluated by using Hyper-Geometric distribution. As suggested by Dyer and Pierce (1993), $p(K)$ is assumed to have Beta-Binomial (Hyper-Binomial) prior distribution

$$p(K) = \frac{N}{K} \frac{\Gamma(K+\alpha)\Gamma(N-K+\beta)}{\Gamma(N+\alpha+\beta)} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{3.3}$$

where Gamma function, $\Gamma(c)$, is defined as $\int_0^\infty e^{-u}u^{c-1}du$. We describe the likelihood of k conditional on K using the Hyper-Geometric distribution:

$$p(k|K) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \tag{3.4}$$

Due to the conjugacy of Beta-Binomial prior probability and Hyper-Geometric likelihood, the posterior probability $p(K|k)$ follows a Beta-Binomial distribution (Grosh (1972)):

$$p(K|k) = \frac{p(k|K)p(K)}{\sum_{K=K_{min}}^{K=K_{max}} p(k|K)p(K)} \tag{3.5}$$

where, we define $K_{min} = \sum_{i=1}^{i=t} k_i$ and $K_{max} = \left(N - \sum_{i=1}^{i=t}(n_i - k_i)\right)$. This can be reorganized as

$$p(K|k) = \binom{N-n}{K-k} \frac{\Gamma(\alpha+K)\Gamma(\beta+N-k)\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+k)\Gamma(\beta+n-k)\Gamma(\alpha+\beta+N)}. \tag{3.6}$$

Due to the conjugacy between $p(k|K)$ and $p(K)$, the distribution $p(K|k)$ at iteration t becomes;

$$p(K|k) = \binom{N - \sum_{i=1}^{i=t} n_i}{K - \sum_{i=1}^{i=t} k_i} \frac{\Gamma(\alpha+K)\Gamma(\beta+N-\sum_{i=1}^{i=t} k_i)\Gamma(\alpha+\beta+\sum_{i=1}^{i=t} n_i)}{\Gamma(\alpha+\sum_{i=1}^{i=t} k_i)\Gamma(\beta+\sum_{i=1}^{i=t}(n_i-k_i))\Gamma(\alpha+\beta+N).} \tag{3.7}$$

as required by Step (3b).

For a particular stratum h, this can be written as $p(K_h|k_h, n_h)$. Furthermore, it is shown in Dyer and Pierce (1993) that the distribution of k given n and $\boldsymbol{\omega}$,

$$p(k|n) = \binom{n}{k} \frac{\Gamma(\alpha+1)\Gamma(\beta+n-k)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+1)}, \tag{3.8}$$

This paper aims to present a concise framework that focuses on claims that either result in fully legitimate or fully overpaid payments, so called all or nothing claims. Particularly, we are interested in $k = 1$ and $k = 0$ in a sample of one draw (unit) ($n = 1$). This leads to

$$p(k = 1|n = 1) = \frac{\alpha}{\alpha+\beta} \text{ and } p(k = 0|n = 1, \alpha, \beta) = 1 - \frac{\alpha}{\alpha+\beta} \tag{3.9}$$

The parameters of $\boldsymbol{\omega}$= $\{\alpha,\ \beta\}$ are updated via the Beta-Binomial model with each observed unit as:

$$\alpha = \alpha_0 + k; \beta = \beta_0 + (n-k), \tag{3.10}$$

where α_0 and β_0 are hyper-prior values which are chosen to be 0.1 for all strata to indicate relative lack of information. The expected information gain of an additional sample, n_f from stratum h, with the outcome k_f, is denoted as $E[\Delta_h](k_f)$. We use k_f and k_{ff} to refer to the number of overpaid claims, k, from future draws of claims, n_f and n_{ff} respectively, given the data and updated parameters. Using the Equation (2.6), $E[\Delta_h](k_f)$ can be written as a double expectation over k_f and k_{ff}:

$$E[\Delta_h(k_f)] = E_{k_{ff}} E_{k_f} log \frac{p(k_{ff}|k, k_f, n_f, h, t)}{p(k_{ff}|k, h, t)}. \tag{3.11}$$

This corresponds to, for stratum h,

$$E[\Delta_h(k_f)] = \int_{k_f} \int_{k_{ff}} p(k_f|k) p(k_{ff}|k, k_f) log(p(k_{ff}|k, k_f)) dk_{ff} dk_f - \int_{k_{ff}} p(k_{ff}|k) log(p(k_{k_{ff}}|k)) dk_{ff} \tag{3.12}$$

as required by Step (3c).

Lastly, Step (3d) requires finding the stratum with highest expected information gain,

$$h^*(t) = argmax_h E[\Delta_h(k_f)] \tag{3.13}$$

The next sample (one or more units) should be drawn from $h^*(t)$.

3.4 Step 4: Estimation

For each stratum, ρ_h is estimated by using the sample proportion of overpaid claims $\hat{\rho_h} = \frac{k_h}{n_h}$. This is unbiased for the case of interest, all or nothing claims where they either result with fully legitimate or fully overpaid submissions.

In line with the governmental guidelines, we aim to be conservative and prevent incorrect recovery demands from the audited provider. Therefore, the 10^{th} percentile of total number of overpaid claims, $K_{h,0.1}$ is retrieved from the posterior distribution of K, Equation (3.7). It is multiplied with the mean payment to obtain the total overpayment recovery for stratum h, $Y_{recovery,h}$ as

$$Y_{recovery,h,Proposed} = K_{h,0.1}\bar{X}_h \tag{3.14}$$

The aggregate total recovery amount, $Y_{recovery,Proposed}$ can be found via:

$$Y_{recovery,Proposed} = \sum_{h=1}^{h=L} Y_{recovery,h,Proposed}. \tag{3.15}$$

3.5 Comparison with Neyman Allocation

This paper includes comparisons of the proposed framework with one of the benchmark methods of practice, the sole use of Neyman Allocation. Particularly, we compare the estimates of proportion of overpaid claims and recovery amounts in addition to computing the average coverage probabilities. Therefore, in the following the estimation procedure used for Neyman Allocation is presented. When we use Neyman Allocation, we allocate all the samples to strata at one iteration. The estimates of ρ_h are computed as $\hat{\rho_h} = \frac{k_h}{n_h}$. We obtain

a total overpayment estimate as in:

$$\hat{Y} = \sum_{h=1}^{h=L} [\hat{\rho}_h \sum_{i=1}^{i=N_h} X_{h,i}] \tag{3.16}$$

Current governmental sampling guidelines (CMS (2001)) aims to be conservative in its demand from the provider. Therefore, in most cases the lower bound of the one-sided 90 percent confidence interval is recommended as the recovery amount from the provider, $Y_{recovery}$. Since we do not consider underpayments, we censure the confidence interval below \$0. For a given sample size adjusted for the finite population size, one sided censured 90 % confidence interval is computed as:

$$Y_{recovery,N.A.} = max(0, \hat{Y} - t_{0.9}^{d.o.f} \sqrt{[\sum_{h=1}^{h=L} (N_h^2 \frac{N_h - n_h}{n_h N_h}) \frac{1}{n_h - 1} \sum_{i=1}^{i=n_h} (y_{h,i} - \bar{y}_h)^2]}) \tag{3.17}$$

where $t_{0.9}^{d.o.f}$ is the 90^{th} percentile of student's t distribution with Satterthwaite's degrees of freedom (d.o.f) as provided in Cochran (2007), Chapter 5.4.

4 Medical Outpatient Claims Data

This paper uses medical claims data with a focus on payment values. We utilize the publicly available *2008 CMS Outpatient Procedures* file (CMS (2010)). We select four procedures that were identified to have frequent overpaid billings in recent investigations (OIG (2012), Youngstrom (2015)). The procedure codes of interest are *J9265, J9310, J0475* and *J9041*, which correspond to injections in surgeries. The resulting data set consist of 8278 claims where the sampling unit is the medical procedure level. We retrieve the actual payment values of the claims. Figure 1 presents the skewed nature of the payment distribution that has multiple local maximums.

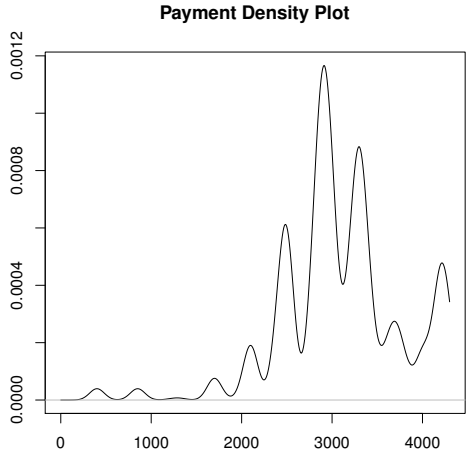

Figure 1: Density plot of payment population

4.1 Claims payment data: Stratification design

In medical audits, the main variable of interest is the overpayment amount which is not known before an audit. Therefore, stratification is generally done using the procedure codes or the payment values with the assumption that the payment values are correlated to the overpayment values. We utilize the second approach in which we stratify the claims using the payment data. Each subsequent stratum consists of increasing dollar amounts with the last stratum consisting of the highest dollar amount of claim payments.

The number of strata is determined as $L = 5$. Next, we choose the stratum boundaries using the approach of Horgan (2006) by using the R package *GA4Stratification* (Keskintürk and Er (2007)). The boundaries of 5 strata are determined as $\{(0-375), [375-1600), [1600-2650), [2650-3550), [3550-4301]\}$. This is how payments are allocated to each stratum and $\boldsymbol{X}_h = \{X_{(1,h)}, ..., X_{(1,N_h)}\}$ for $h = 1, ..., 5$ is determined.

The goal of a stratification scheme used for sampling is to create groups of data that are internally homogenous but as different from each other as possible. Table 1 presents the descriptive statistics of payment values which has been successfully stratified.

Stratum	Mean	Sd	2.5%	25%	Median	75%	97.5%	N_s
s=1	69.40	46.92	20.00	30.00	40.00	40.00	50.00	3949
s=2	915.69	180.87	50.00	50.00	70.00	90.00	100.00	1402
s=3	2335.37	244.61	1700.00	2100.00	2500.00	2500.00	2500.00	588
s=4	3076.60	206.76	2800.00	2900.00	3000.00	3300.00	3400.00	1675
s=5	4012.65	246.71	3600.00	3700.00	4100.00	4200.00	4300.00	664
Overall	1298.47	1441.22	30.00	50.00	600.00	2900.00	4200.00	N=8278

Table 1: Descriptive statistics of payment

4.2 Generation of overpayment scenarios

In order to demonstrate the versatility of the proposed framework, it should be tested for different overpayment patterns. However, the availability of overpayment data is limited among the publicly available resources and publications, mainly due to confidentiality concerns as well as its cost. It should be noted that an audit needs to be conducted to retrieve overpayment data. Therefore, in order to demonstrate the versatility of our framework, we resort to simulation to construct a number of overpayment scenarios. This is how we retrieve $\boldsymbol{Y_h}$ vectors for each stratum, $h = 1, ..., L$. In the following, we describe the details and the choice of parameters for data generation.

First, the probability matrix $\boldsymbol{\pi}$ is defined to list the proportion of overpaid claims in each stratum. Particularly, we generate the overpayment data for the $i^{th}(i = 1, ..., 7)$ scenario and h^{th} stratum via the overpayment proportion, $\pi_{i,h}$. 7 scenarios are used to represent a broad but manageable number set of cases, see Table 2. It should be emphasized that we make the assumption of "all or nothing" claims. This corresponds to all claims being either fully legitimate or overpaid. In other words, we assume there are no partial overpayments.

For instance, in the first overpayment scenario, the first stratum consists of fully legitimate claims where the proportion of overpaid claims, $\pi_{1,1}$ is 0. Whereas all the claims in the fifth stratum are fraudulent as indicated by $\pi_{1,5} = 1$. Scenarios 1 and 2 represent cases that have different proportions of overpaid claims in each stratum. Scenario 3 and Scenario 4 are modified versions in that the overpayment proportions in stratum 1 and 5 are modified. Scenarios 5, 6 and 7 have different, but fixed proportions of overpaid claims throughout all

Scenario	$\pi_{i,h}$	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$
i=1	$\pi_{1,h}$	0.00	0.25	0.50	0.75	1.00
i=2	$\pi_{2,h}$	1.00	0.75	0.50	0.25	0.00
i=3	$\pi_{3,h}$	0.05	0.25	0.50	0.75	0.95
i=4	$\pi_{4,h}$	0.95	0.75	0.50	0.25	0.05
i=5	$\pi_{5,h}$	0.50	0.50	0.50	0.50	0.50
i=6	$\pi_{6,h}$	0.25	0.25	0.25	0.25	0.25
i=7	$\pi_{7,h}$	0.75	0.75	0.75	0.75	0.75

Table 2: Probability of overpayment proportions in each stratum per scenario

strata.

Using this choice of parameters, we run a simulation of size 1000, and record the overpayment data. In order to provide a better understanding of the overpayment data, we report the descriptive statistics of a particular scenario. Table 3 presents the descriptive statistics of the 1000 generations of overpayment data for Scenario 1.

Stratum	Mean	Sd	2.5%	25%	Median	75%	97.5%
s= 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
s= 2	228.92	406.58	0.00	0.00	0.900	257.97	1009.29
s= 3	1167.69	1180.43	0.00	0.00	860.72	2492.32	2500.00
s= 4	2307.45	1344.07	0.00	1421.13	2900.00	3228.89	3400.00
s= 5	4012.65	246.89	3600.00	3700.00	4100.00	4200.00	4300.00
Overall	910.47	1464.00	0.00	0.00	0.00	2469.77	4200.00

Table 3: Descriptive statistics of overpayment for Scenario 1

The overpayment values in each stratum are well separated from each other in relation to their average values. The larger variance values of overpayments correspond to greater uncertainty. Hence the greatest uncertainty can be argued to exist in the fourth stratum.

5 Illustration and Analysis

This section presents an illustration of the proposed method and discussion on its validity and efficiency. It also provides comparisons with Neyman Allocation for different overpayment scenarios and allocation settings.

Firstly, we demonstrate the sampling allocation at each iteration for overpayment scenario 1 for a randomly chosen replication using 45 initial and 10 additional sample sizes. We demonstrate this overpayment scenario since it includes strata with totally legitimate or overpaid claims as well as strata with varying proportion of overpaid claims.

We allocate the initial set of 45 units to 5 strata using Neyman Allocation at $t = 0$ and then use the proposed information-theoretic framework for 10 additional units. At each iteration, a claim is sampled from the stratum with the highest expected information gain. We report the number of claims sampled in each stratum at particular iterations in Table 4. The numbers within parentheses represent the overpaid claims (where $y = x$, $k = 1$).

	s=1	s=2	s=3	s=4	s=5
$t = 0$	8 (0)	10 (1)	6 (2)	14 (11)	7 (7)
$t = 1$	0	0	1 (1)	0	0
$t = 2$	0	0	1 (1)	0	0
$t = 3$	0	0	1 (1)	0	0
$t = 4$	0	1 (0)	0	0	0
$t = 5$	0	1 (0)	0	0	0
$t = 6$	0	1 (0)	0	0	0
$t = 7$	0	1 (0)	0	0	0
$t = 8$	0	1 (0)	0	0	0
$t = 9$	0	1 (0)	0	0	0
$t = 10$	0	1 (1)	0	0	0
Total	8(0)	17(2)	9 (5)	14 (11)	7(7)

Table 4: Sample allocation of 45+10 claims of each stratum for overpayment scenario 1

Initial Neyman allocation leads to 8 and 7 claims inspected in stratum 1 and stratum 5, respectively. These are found to be either fully legitimate claims as in stratum 1, or fully overpaid as in stratum 5. Therefore, the expected gain of information from observing an additional unit (sample of one) is minimal and no more claims are sampled from these strata in the following iterations. Entropy is expected to be higher when the proportion of overpaid claims in the sample of a particular stratum is close to 0.5. As expected, the highest uncertainty and expected information gain occur at stratum 3. Samples of overpaid claims gradually decrease the uncertainty and expected information gain for their respective stratum. This eventually changes the stratum to sample from, and stratum 2 is allocated the

remainder of the resources starting at $t = 4$. In contrast, if we used the Neyman Allocation for the whole set of 55 samples, we would have sampled $9, 13, 7, 18, 8$ claims from stratum 1 through stratum 5 respectively.

Next, we investigate the versatility of the proposed approach for a number of overpayment scenarios and resource allocation designs. Table 5 presents a number of resource allocation designs. Different resource allocation designs allow us to investigate the performance trade-offs for various settings of the initial sample sizes and additional sample sizes. In the first step, a group of samples are allocated to strata via Neyman Allocation, providing us with the initial information about overpayments. The remainder of the samples are allocated one at a time at each iteration to the stratum with the highest amount of expected information gain.

Allocation Strategy	Initial	Additional	Total
A_1	45	10	55
A_2		20	65
A_3		30	75
A_4	60	10	70
A_5		20	80
A_6		30	90
A_7	90	10	100
A_8		20	110
A_9		30	120

Table 5: Resource Allocation Designs

Next, we compare the estimation errors of the proportion of overpaid claims of each stratum for both Neyman Allocation and the proposed method for resource allocation strategies and refer to them as $\hat{\rho}_{hProposed}$ and $\hat{\rho}_{hN.A.}$. Using 1000 replications of 7 overpayment scenarios, we compute the overall mean absolute error of the proportion of overpaid claims, MAE_ρ by using the relative stratum sizes as weights:

$$MAE_\rho = \sum_{h=1}^{h=L} \frac{N_h}{N} |\hat{\rho}_h - \rho_h|.$$

Table 6 lists the MAE_ρ values of both methods for 7 overpayment scenarios and 9 al-

A	Total Sample Size	MAE_ρ	O.S. 1	O.S. 2	O.S. 3	O.S. 4	O.S. 5	O.S. 6	O.S. 7
A_1	55	Proposed	3.26%	8.25%	7.91%	8.69%	11.72%	10.99%	11.45%
	55	N.A.	3.39%	9.95%	8.76%	8.77%	11.53%	11.11%	12.00%
A_2	65	Proposed	2.91%	7.13%	7.28%	7.97%	11.09%	10.06%	10.19%
	65	N.A	2.97%	9.08%	7.87%	7.95%	10.68%	10.08%	10.32%
A_3	75	Proposed	2.78%	6.85%	6.80%	7.41%	10.37%	9.58%	9.30%
	75	N.A	2.77%	8.15%	7.22%	7.41%	9.67%	9.60%	10.02%
A_4	70	Proposed	2.69%	7.02%	7.12%	7.61%	10.25%	9.79%	9.60%
	70	N.A	2.92%	8.46%	7.60%	7.50%	10.08%	9.68%	9.94%
A_5	80	Proposed	2.68%	6.37%	6.60%	7.10%	10.05%	8.75%	9.07%
	80	N.A	2.74%	8.11%	7.01%	7.08%	9.51%	8.96%	9.85%
A_6	90	Proposed	2.49%	6.05%	6.04%	6.49%	9.38%	8.27%	8.48%
	90	N.A	2.53%	7.37%	6.53%	6.61%	8.93%	8.44%	8.89%
A_7	100	Proposed	2.31%	5.95%	5.98%	6.33%	8.50%	8.19%	8.24%
	100	N.A	2.44%	7.09%	6.31%	6.34%	8.34%	8.25%	8.62%
A_8	110	Proposed	2.17%	5.61%	5.66%	6.13%	8.16%	7.58%	7.69%
	110	N.A	2.29%	6.86%	5.99%	5.92%	8.07%	7.75%	8.47%
A_9	120	Proposed	2.12%	5.38%	5.41%	5.71%	8.15%	7.25%	7.44%
	120	N.A	2.20%	6.53%	5.55%	5.52%	7.78%	7.45%	7.74%

Table 6: MAE_ρ for both methods for each overpayment scenario (OS) and sample allocation (A)

location designs. It can be argued that in general, the proposed method results in lower estimation error for the proportion of overpaid claims in the population. When the overpayment probability is different across strata, the proposed method can allocate audit resources relatively more efficiently via learning about strata overpayment probability distributions. Overpayment scenario 5 (O.S. 5) is the only case where the proposed method consistently performs worse. As Table 2 presents, this is the scenario where the probability of overpayment is 0.5 through all strata. This corresponds to the case with the highest uncertainty over a Bernoulli event if the sample claim is legitimate or overpaid. Therefore, the entropy method does not provide any advantages over the Neyman Allocation method.

Our proposed method performs well for a large enough initial sample size and provides better overpayment estimates. Furthermore, higher additional sample size to utilize the initial information improves the performance of the proposed method. Figure 2 presents the boxplots of differences of MAE_ρ for both methods using 1000 replications. It can be

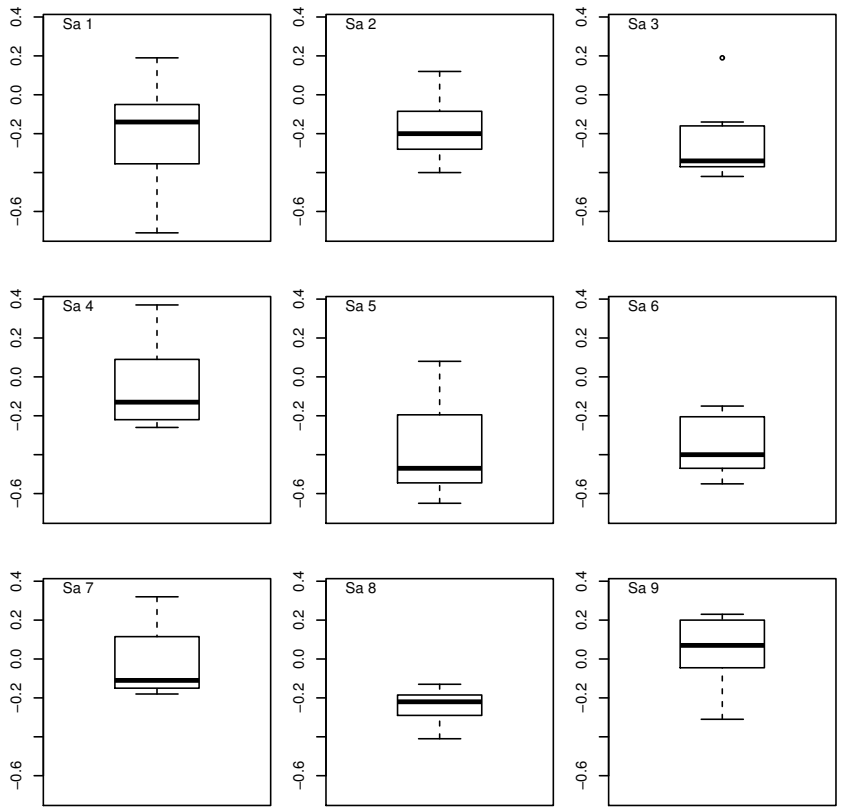

Figure 2: Boxplots of $MAE_{\rho,Proposed} - MAE_{\rho,N.A.}$ for overpayment scenario 1 and all allocation strategies

seen that the advantage of the proposed method increases when the additional sample size becomes larger. For instance, the first 3 sample allocations A_1, A_2 and A_3 show the increasing performance of the proposed entropy based method relative to the Neyman Allocation method. We can also argue that the relative advantage of the entropy method is the smallest in the cases of small additional sample sizes. In addition, Figure 2 reveals that the advantage of the proposed method decreases when the total sample size increases. This is not surprising since the increase in sample size can be expected to decrease the incremental benefit of efficient allocation of resources. However, it should also be noted that the total sample size is generally small for medical audits, thus making the allocation of resources critical.

We conduct a one-directional paired samples t-test to evaluate the significance of these

mean absolute errors. As it can be seen from Table 7, the p-values indicate the superior performance of the proposed method over the Neyman Allocation method. In 56 cases among the 72 hypothesis tests, we ascertain evidence for the alternative hypothesis that the proposed method has a lower Mean Absolute Error compared to the Neyman Allocation method when the level of significance is 0.05. Overall, the differences are shown to be statistically significant, in addition to being practically significant, as discussed.

Allocation Strategy (A)	O.S. 1	O.S. 2	O.S. 3	O.S. 4	O.S. 5	O.S. 6	O.S. 7
A_1	0.00	0.00	0.00	0.00	0.14	0.74	0.85
A_2	0.00	0.00	0.00	0.00	0.18	0.52	0.95
A_3	0.00	0.00	0.00	0.00	0.00	0.37	0.01
A_4	0.00	0.00	0.00	0.00	0.44	0.59	0.67
A_5	0.00	0.00	0.00	0.00	0.35	0.00	0.02
A_6	0.00	0.00	0.00	0.00	0.00	0.33	0.14
A_7	0.00	0.00	0.00	0.00	0.32	0.00	0.02
A_8	0.00	0.00	0.00	0.01	0.00	0.00	0.00
A_9	0.00	0.00	0.01	0.02	0.33	0.03	0.73

Table 7: p values for the paired sample t-tests with $H_a : \mu_{MAE_{\rho,N.A.}-MAE_{\rho,Proposed}} \geq 0$

In order to investigate the validity of the proposed model, we compute average coverage probabilities. These are evaluated by finding the lower bound value of the 90% confidence interval using sample statistics and comparing it against the actual mean overpayment value. Using 1000 replications of 7 overpayment scenarios, we compute the percentage of times the actual overpayment value is higher than the computed lower bound value, as reported by Table 8. Both methods generally provide reasonable coverage with the exception of overpayment scenarios of 5 and 7. Both methods are still comparable, and even for overpayment scenario 7, the coverage of the proposed method does not fall below 82.30%.

Lastly, we compare the recovery amounts of both methods with the actual overpayment amount for each allocation strategy and overpayment scenario. For the Neyman Allocation method, we use the standard strategy of computing the average recovery amount as the lower bound of 90% confidence interval via Equation (3.17). For the proposed method, the probabilistic framework is utilized via Equation (3.14). Table 9 presents the difference of errors in estimating overpayment, $MAPE[Y_{recovery,N.A.}] - MAPE[Y_{recovery,Proposed}]$. As

A	Method	O.S. 1	O.S. 2	O.S. 3	O.S. 4	O.S. 5	O.S. 6	O.S. 7
A_1	Proposed	100.00%	99.40%	97.30%	99.30%	95.70%	95.90%	90.70%
	N.A	100.00%	99.90%	98.00%	98.90%	88.40%	91.10%	83.00%
A_2	Proposed	99.90%	99.50%	97.90%	98.20%	87.60%	88.60%	85.70%
	N.A	99.90%	99.90%	97.30%	99.10%	88.20%	89.70%	86.10%
A_3	Proposed	100.00%	99.50%	98.10%	97.50%	87.70%	91.60%	83.20%
	N.A	100.00%	99.90%	98.20%	98.80%	88.00%	90.50%	85.50%
A_4	Proposed	100.00%	99.80%	98.00%	98.10%	86.90%	90.20%	83.60%
	N.A	100.00%	99.80%	98.60%	98.30%	88.30%	91.10%	85.30%
A_5	Proposed	100.00%	99.50%	97.60%	98.00%	88.20%	90.70%	82.30%
	N.A	100.00%	99.60%	97.60%	98.20%	88.00%	89.90%	84.00%
A_6	Proposed	99.80%	99.70%	97.90%	97.60%	86.30%	90.30%	86.80%
	N.A	99.90%	99.90%	97.80%	98.00%	87.20%	90.00%	87.30%
A_7	Proposed	99.80%	99.60%	96.20%	98.90%	90.40%	90.10%	85.50%
	N.A	99.80%	99.70%	97.20%	98.50%	90.20%	89.90%	84.60%
A_8	Proposed	100.00%	99.80%	97.70%	98.20%	88.10%	89.40%	85.80%
	N.A	99.90%	99.80%	97.90%	98.50%	88.50%	90.50%	86.80%
A_9	Proposed	99.80%	99.80%	97.90%	98.50%	87.30%	87.80%	86.60%
	N.A	99.70%	99.80%	97.50%	98.80%	88.00%	88.70%	87.60%

Table 8: Average coverage probabilities for the Proposed and N.A methods

A	OS_1	OS_2	OS_3	OS_4	OS_5	OS_6	OS_7
A_1	4.64%	41.46%	4.75%	34.13%	-0.15%	-2.50%	9.82%
A_2	4.21%	39.13%	4.01%	33.30%	-2.28%	-2.74%	6.37%
A_3	4.44%	35.52%	3.90%	29.57%	-3.46%	-6.06%	8.35%
A_4	4.19%	28.27%	1.93%	23.85%	-2.75%	-6.18%	3.61%
A_5	3.49%	26.88%	2.17%	20.97%	-3.25%	-6.74%	3.45%
A_6	2.96%	26.37%	0.57%	19.69%	-4.99%	-6.73%	1.94%
A_7	3.24%	23.82%	0.83%	18.77%	-3.24%	-5.24%	1.97%
A_8	3.40%	22.72%	0.93%	17.74%	-2.97%	-5.38%	1.54%
A_9	2.77%	20.98%	0.70%	16.35%	-3.05%	-8.45%	1.53%

Table 9: $MAPE[Y_{recovery,N.A.}] - MAPE[Y_{recovery,Proposed}]$., Mean Absolute Percentage Error Differences of Recovery Amounts

expected, the performance of both methods increase with increasing sample sizes. For all overpayment scenarios except 5 and 6, the proposed method results in better estimation of recovery amount compared to the Neyman Allocation process. The lack of superior performance for those scenarios can be explained by the poor estimation of 10^{th} percentile of K because of the relatively lower number of overpayments. A potential improvement can

be to use a multivariate distribution for K to allow pooling of information from S strata. For all other scenarios, Neyman improves itself when the strata with higher payments have higher probabilities of claims being overpayments. The proposed method is neutral to those changes, and still outperforms Neyman Allocation.

6 Conclusion

This paper presents an information theoretic iterative sampling framework to estimate the proportion and number of overpaid claims as well as the overpayment recovery amount. We utilize Lindley's entropy measure to evaluate the expected amount of information from the prospective samples. The proposed framework is shown to provide valid and efficient overpayment recovery estimates for an all or nothing claims. Moreover, it outperforms the Neyman Allocation method in terms of estimation error for a variety of allocation strategies and overpayment scenarios. The efficiency gains are highest when the overpayment probabilities are different across strata. In addition, it provides the probability distributions of the number of overpaid claims in a given stratum which, in turn, offers better estimates. The evaluation of information of the samples can be especially beneficial for outpatient procedures with high medical audit costs. Moreover, the proposed method is general; thus, it can be adapted to settings that can benefit from an iterative stratified sampling framework.

The proposed method can be modified further for improved versatility. It should be noted that, as it is, the proposed method only uses the information content to make iterative allocations. Weighted combinations of the information value and payment amounts can be beneficial. To determine the weights, a decision model similar to Fouskakis et al. (2009) can incorporate the real life costs and benefits using the auditor expertise. Ignatova and Edwards (2008) provide a related discussion for the incorporation of the costs within a multi-stage sampling framework. Another implementation consideration involves the trade-off between the initial sample size and additional sample size. One should have a large enough initial sample size to start with good estimates, and enough number of additional samples for learning. An optimization model can be proposed for optimal determination of the initial

and additional number of samples.

A potential extension could be to use a sampling framework with overpayment models that allow partial overpayments. This would require building more complicated and non-conjugate posterior distributions. These can numerically be computed using Markov chain Monte Carlo simulation methods. Furthermore, we have assumed a lack of knowledge and used vague uninformative priors within the proposed Bayesian framework. This helps with creating a fair assessment of the provider under investigation. One can argue that potential auditor knowledge can be used to construct informative priors which may result in better sampling allocation decisions. On the other hand, Edwards et al. (2015) argue that Bayesian approaches may be abused when the provider might be asked for recoupment although the sample shows no evidence of impropriety. That may be a cause of concern in case of using very strong priors with a small sample. We should emphasize that relative lack of information is assumed in this manuscript, therefore that should not be a concern.

References

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304.

CMS (2001). Program memorandum carriers transmittal b-01-01. `http://www.cms.gov/Regulations-and-Guidance/Guidance/Transmittals/downloads/B0101.pdf`. Accessed: 03/03/2015.

CMS (2010). Basic Stand Alone (BSA) Medicare claims public use files (PUFs). `https://www.cms.gov/bsapufs`. Accessed: 11/07/2016.

CMS (2017). NHE Fact Sheet. The Centers for Medicare & Medicaid Services. `https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html`. Accessed: 07/01/2017.

Cochran, W. G. (2007). *Sampling Techniques*. John Wiley & Sons.

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

Dyer, D. and Pierce, R. L. (1993). On the choice of the prior distribution in Hypergeometric sampling. *Communications in Statistics - Theory and Methods*, 22(8):2125–2146.

Ebrahimi, N., Soofi, E., and Soyer, R. (2010). Information measures in perspective. *International Statistical Review*, 78(3):383412.

Edwards, D., Gilliland, D., Ward-Besser, G., and Lasecki, J. (2015). Conservative penny sampling. *Journal of Survey Statistics and Methodology*, 3(4):504–523.

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

Ekin, T., Ieva, F., Ruggeri, F., and Soyer, R. (2018). Statistical methods for medical fraud assessment: Exposition to an emerging field. *International Statistical Review*.

Ekin, T., Musal, R. M., and Fulton, L. V. (2015). Overpayment models for medical audits: multiple scenarios. *Journal of Applied Statistics*, 42(11):2391–2405.

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009). Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *The Annals of Applied Statistics*, 3(2):663–690.

Grosh, D. L. (1972). A Bayes sampling allocation scheme for stratified finite populations with Hyperbinomial prior distributions. *Technometrics*, 14(3):599–612.

Guthrie, D. et al. (1989). Statistical models and analysis in auditing: Panel on nonstandard mixtures of distributions. *Statistical Science*, 4:2–33.

Hidiroglou, M. A. and Kozak, M. (2017). Stratification of skewed populations: A comparison of optimisation-based versus approximate methods. *International Statistical Review*.

Horgan, J. M. (2006). Stratification of skewed populations: A review. *International Statistical Review*, 74(1):67–76.

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

Jeffries, M. (2016). Outpatient procedures. `http://health.howstuffworks.com/health-insurance/outpatient-procedure1.htm`. Accessed: 11/07/2016.

Keskintürk, T. and Er, Ş. (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics & Data Analysis*, 52(1):53–67.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5):797–806.

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

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005.

Mathew, O. O., Sola, A. F., Oladiran, B. H., and Amos, A. A. (2014). Efficiency of Neyman allocation procedure over other allocation procedures in stratified random sampling. *American Journal of Theoretical and Applied Statistics*, 2(5):122–127.

Musal, R. M. and Ekin, T. (2017). Medical overpayment estimation: A Bayesian approach. *Statistical Modelling*, 17(3):196–222.

Musal, R. M. and Soyer, R. (2014). Estimation of group priorities and value of information. *Journal of Multi-Criteria Decision Analysis*, 21(3-4):173–181.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

OIG (2012). Review of Medicare outpatient billing for selected drugs at Essentia Health Duluth. `https://oig.hhs.gov/oas/\reports/region9/91202021.pdf`. Accessed:03/06/2015.

Park, R. and Perling, L. (2016). Statistical sampling: Evolving legal issues. `https://www.healthlawyers.org/Events/Programs/Materials/Documents/MM12/papers/D_park_perling.pdf`. Accessed: 04/30/2016.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(4):623 – 656.

Shewry, M. C. and Wynn, H. P. (1987). Maximum entropy sampling. *Journal of applied statistics*, 14(2):165–170.

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

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

Soofi, E. S. (2000). Principal information theoretic approaches. *Journal of the American Statistical Association*, 95(452):1349–1353.

Youngstrom, N. (2015). Medical-necessity audits gain steam, hit on chemo, cardiac; watch for new lcds. `http://www.racmonitor.com/rac-enews/1833-medical-necessity-audits-gain-steam-hit-on-chemo-cardiac-watch-for-new-lcds.html`. Accessed: 11/07/2016.