Health Insurance Market Risk Assessment: Covariate Shift and k-Anonymity

Dennis Wei* Karthikeyan Natesan Ramamurthy* Kush R. Varshney*

Abstract

Health insurance companies prefer to enter new markets in which individuals likely to enroll in their plans have a low annual cost. When deciding which new markets to enter, health cost data for the new markets is unavailable to them, but health cost data for their own enrolled members is available. To address the problem of assessing risk in new markets, i.e., estimating the cost of likely enrollees, we pose a regression problem with demographic data as predictors combined with a novel three-population covariate shift. Since this application deals with health data that is protected by privacy laws, we cannot use the raw data of the insurance company's members directly for training the regression and covariate shift. Therefore, to construct a full solution, we also develop a novel method to achieve k-anonymity with the workload-driven quality of data distribution preservation achieved through dithered quantization and Rosenblatt's transformation. We illustrate the efficacy of the solution using real-world, publicly available data.

1 Introduction.

The Patient Protection and Affordable Care Act, known colloquially as Obamacare, changed the landscape of health insurance in the United States significantly. Health insurance companies entered new markets, defined by geography, by age group, and by other prospect base criteria. When the legislation was being enacted, the companies had to decide which new markets to enter using the information at their disposal at the time. In making the decisions, companies sought to enter markets containing an abundance of profitable, i.e. low-cost, individuals likely to enroll in their plans. (Although we undertook the work presented in this paper motivated by Obamacare, the desideratum to enter markets with low-cost individuals is always true, regardless of whether there has been a significant change in the landscape.)

Healthcare cost and utilization data has been analyzed for a variety of public health-related and health economics-related tasks, usually through regression techniques that attempt to take the lack of normality and the heteroscedasticity of costs into account [8, 6, 26]; health cost and utilization data often follows a distribution similar to the delta-lognormal distribution [8, 4]. Ordinary least-squares regression with and without log-transformed data, two-part models, generalized linear models, and multiplicative regression have all been used successfully to predict the healthcare costs of individuals. However, it should be noted that we have learned through conversations with health insurance companies that, to-date, only crude, inaccurate models have been applied for assessing market risk; sophisticated regression models have not yet been used.

Critically, although insurers have health cost data for their own members available at decision-making time, health cost data for individuals in new markets is not available to them for a variety of reasons. This situation calls for predictive modeling to estimate health cost profiles of enrollees in the new markets, which can be appropriately summarized into risk statistics for the new markets. In this problem, demographic data for current markets and new markets is available from public sources. Therefore, in this work, we develop a predictive analytics approach in which we estimate the relationship between demographics and costs in the current member population and then apply the learned model to the new market's demographic data, taking into account the difference between the demographic distribution of the current member population and the demographic distribution of the prospective enrollees in the new market. This setting is known as *covariate shift* in the machine learning literature and has been studied for the types of regression models used with health cost data [23, 19].

Specifically for the new market risk assessment problem, since only a subset of individuals in a market enroll in a health insurance company's plans, we have an additional population to consider beyond simply the individuals in the old and new markets. Existing methods dealing with covariate shift can be cateogorized as *two-population* methods; in this work, the problem of interest requires a *three-population* shift, which has not appeared in the literature. The first main contribution of this paper is the development of three-population shift methods for healthcare market risk assessment.

Health data on individuals in the United States, even internal use by an insurance company for its plan-

*Mathematical Sciences and Analytics Department, IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA.

ning and strategy, is protected by the Health Insurance Portability and Accountability Act. Privacy of individuals must not be compromised. Several statistical interpretations of the legal language for protecting privacy exist; the property of k-anonymity is a common interpretation [17]. Under the k-anonymity privacy model, the data for an individual cannot be distinguished from at least $k-1$ other individuals [25, 22].

Given an initial data set, k-anonymity is often achieved using generalizations and suppressions [10]. It is known that multidimensional generalization has superior performance than single-dimensional generalization [12], and treating anonymization as a clustering problem offers more flexibility than using predefined generalization hierarchies [7]. To achieve k-anonymity, we would like to group the samples or records in the data by similarity such that the smallest group has at least k elements. Any such grouping or clustering is equally good from the privacy perspective; it is the *workload* for which the data is to be used that defines the quality of the grouping [22, 10, 27]. In our case, the workload is three-population shift-based market risk prediction.

Most existing clustering algorithms such as k-means clustering take the number of clusters as an input parameter rather than the minimum number of samples in each cluster, which is what is needed for k-anonymity. The k-member clustering problem, which has the k of k-anonymity as the input parameter rather than the number of clusters, and its solution with a greedy algorithm is proposed in [7]. This problem is given the name probability-constrained quantization by [20] and a modification of the k-means Lloyd-Max algorithm is developed for its solution. This problem is also related to maximum output entropy quantization [18].

In these clustering approaches to anonymization, the optimization criterion is based on the average distance or distortion of the individual samples. With such a criterion, the optimal representation points, i.e. cluster centers, do not follow the same distribution as the original data. (Without the probability constraint, the optimal quantizer point density, which is the distribution of the cluster center locations in the asymptotic limit as the number of clusters goes to infinity is a normalized version of the original data distribution to the one third power [9].) We raise this point of the distribution of representation points not following the distribution of the data because the distribution of the data has an important role in the workload of our interest: regression with covariate shift.

Distribution-preserving quantization is an alternative method to the standard k-means or standard quantization approaches that has the desired representation point behavior and has never been considered in the privacy preservation context before [13, 5]. The approach of [13] is based on subtractive dithered quantization [15] followed by Rosenblatt's transformation [21]. Dithering, the introduction of noise or random perturbations, is a technique for privacy preservation fraught with several issues [11], but in our work, the introduction of noise is not for the purpose of privacy preservation, but to allow the manipulation of the distribution.

Existing approaches for distribution-preserving quantization take the number of clusters as a parameter, just like the standard clustering and quantization methods, but which is not amenable to achieving good k-anonymity. To the best of our knowledge, there is no existing probability-constrained, density-preserving quantization algorithm, which is what is required for k-anonymization followed by three-population shift-based prediction for healthcare market risk assessment. The second main contribution of this paper is the development of such a method.

We apply the proposed methods to publicly-available health data from the Medical Expenditure Panel Survey (MEPS) produced by the United States Department of Health and Human Services' Agency for Healthcare Research and Quality. We demonstrate the efficacy of our solution, modeling the entire MEPS data set as the existing market and individual rating areas in California as the new markets, and modeling plan enrollment using true Obamacare enrollment distributions [2, 3]. The proposed three-population shift significantly improves the aggregate prediction accuracy and the proposed privacy-preservation algorithm does not degrade prediction accuracy much.

The remainder of the paper is organized as follows. In Section 2, we provide background on the covariate shift problem. In Section 3, we detail the exact problem statement required for new market risk assessment and propose approaches for its solution. Next, to make risk assessment tenable in the healthcare domain, we develop a new privacy-preservation method for the market risk assessment workload that combines aspects of k-member clustering and distribution-preserving quantization in Section 4. We present empirical results on real-world healthcare data in Section 5. Section 6 provides a summary and discussion.

2 Background on Covariate Shift Problem.

In this section, we first introduce notation and then describe the basic learning problem encountered in the covariate shift setting. This section is very general; we specialize the exposition to health cost data in Section 3.

Consider the following problem: we wish to predict a response variable Y using predictor variables X. Given a class of functions $\mathcal{F}$ and training samples

(x_i, y_i), $i = 1, \ldots, n$, a predictor function is selected from $\mathcal{F}$ to minimize the empirical risk,

$$\hat{Y}(\cdot) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i), \tag{2.1}$$

for some choice of loss function $\mathcal{L}$ that measures the error between the predicted response $f(x_i)$ and actual response y_i.

Assume that the training samples are drawn i.i.d. from the joint distribution $p_{X,Y} = p_X p_{Y|X}$. The problem of *covariate shift* occurs when the predictor variables or covariates are drawn from a different distribution q_X in the test phase. The conditional distribution $p_{Y|X}$ is assumed to remain the same. As the number of samples $n \to \infty$, the empirical risk in (2.1) converges to the population risk

$$\mathbb{E}\left[\mathcal{L}(f(X), Y)\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{L}(f(X), Y) \mid X\right]\right],$$

from which it can be seen that the optimal choice of predictor f depends only on the conditional distribution $p_{Y|X}$, regardless of the marginal distribution for X (e.g. p_X or q_X). Hence as $n \to \infty$, the conditional distribution $p_{Y|X}$ can be learned very accurately and the optimal predictor can be obtained provided that the class $\mathcal{F}$ is rich enough to contain it. However, when n is finite and/or $\mathcal{F}$ is overly constrained, then the predictor $\hat{Y}$ resulting from (2.1) generally depends on the training distribution p_X and thus can be mismatched to the test distribution q_X under which performance is evaluated.

A straightforward solution to covariate shift is to weight the training samples by the ratio $q_X(x_i)/p_X(x_i)$. This weighting represents the relative importance of each sample under q_X rather than p_X. The weighted empirical risk

$$\frac{1}{n} \sum_{i=1}^{n} \frac{q_X(x_i)}{p_X(x_i)} \mathcal{L}(f(x_i), y_i)$$

then converges to

$$\mathbb{E}_{p_X p_{Y|X}} \left[\frac{q_X(X)}{p_X(X)} \mathcal{L}(f(X), Y)\right] = \mathbb{E}_{q_X p_{Y|X}} \left[\mathcal{L}(f(X), Y)\right], \tag{2.2}$$

thus matching the test distribution.

In practice, the distributions p_X, q_X and in particular their ratio need to be estimated from data. We assume for simplicity and in accordance with the market shift application that the predictor variables X are discrete, taking values in a set $\mathcal{X}$. Then the probability mass functions (PMFs) of interest can be approximated by the empirical distributions $\hat{p}_X(x)$, $\hat{q}_X(x)$ and their ratio by $\hat{q}_X(x)/\hat{p}_X(x)$. Rewriting the weighted empirical risk as an outer sum over $\mathcal{X}$ and an inner sum over training samples with common $x_i = x$, we have

$$\sum_{x:\hat{p}_X(x)>0} \hat{p}_X(x) \frac{\hat{q}_X(x)}{\hat{p}_X(x)} \frac{1}{n(x)} \sum_{i:x_i=x} \mathcal{L}(f(x), y_i) = \sum_{x:\hat{p}_X(x)>0} \hat{q}_X(x) \frac{1}{n(x)} \sum_{i:x_i=x} \mathcal{L}(f(x), y_i), \tag{2.3}$$

where $n(x)$ is the number of training samples with $x_i = x$. As desired, under the assumption that the support of $\hat{p}_X$ asymptotically contains the support of $\hat{q}_X$, the weighted empirical risk (2.3) converges to (2.2) as $n \to \infty$.

The non-parametric empirical distribution approach discussed above appears to work well when X is discrete and the number of possible values $|\mathcal{X}|$ is not too large compared to the sample size. The latter condition is satisfied if the number of predictor variables is small and the number of possible values for each variable is also modest. However for large $|\mathcal{X}|$ or continuous X, estimating $p_X(x)$, $q_X(x)$ and/or their ratio becomes difficult and a parametric form may need to be assumed. In Section 3.1, we discuss one such parametric approximation involving logistic regression.

3 Market Risk Assessment.

The covariate shift framework in Section 2 can be applied to health care cost analysis for market risk assessment. In particular, the response variable Y of interest is the annual cost of a member to the insurance company. The predictor variables are demographic features such as age, gender, income, veteran status, smoking status, place of residence, place of origin, and so on. The response variable being continuous, the learning problem (2.1) is one of regression. For example, the set $\mathcal{F}$ may be the set of all linear predictors with an ℓ_1-norm less than a fixed constant.

There are some nuances to consider beyond the general covariate shift problem recapitulated in Section 2 in our application of interest. Notably, there is a distinction between member populations and larger market populations from which member populations are enrolled. To denote the different populations, we use the binary variables E and M. The variable E indicates enrollment in an insurance company's plan ($E = 1$ means enrolled), and the variable M differentiates the existing current market from the new market ($M = 1$ means new market).

Training data with costs is assumed to come from the insurance company's data on current plan members. Thus the training distribution p_X in Section 2 is now denoted as $p_{X|E,M}(x \mid e = 1, m = 0)$, referring to enrollees in the current market. Likewise, the test distribution

q_X is $p_{X|E,M}(x \mid e = 1, m = 1)$ for enrollees in the new market. In some cases it may be possible to obtain $p_{X|E,M}(x \mid 1, 1)$ directly if enough is known about potential enrollees in the new market. If this is true then the basic covariate shift framework in Section 2 applies. We refer to this case as the *two-population* market shift problem.

The more general case is that $p_{X|E,M}(x \mid 1, 1)$ cannot be estimated directly, which warrants the consideration of a *three-population* version of the problem. Besides the current member distribution $p_{X|E,M}(x \mid 1, 0)$, it is assumed that demographic distributions for the current and new markets are available, corresponding to $p_{X|M}(x \mid 0)$ and $p_{X|M}(x \mid 1)$ respectively. These distributions are related by Bayes' rule,

$$p_{X|E,M}(x \mid 1, m) = \frac{p_{E|X,M}(1 \mid x, m) p_{X|M}(x \mid m)}{p_{E|M}(1 \mid m)}, \quad m = 0, 1. \tag{3.4}$$

Taking the ratio of $m = 1$ to $m = 0$ gives

$$\frac{p_{X|E,M}(x \mid 1, 1)}{p_{X|E,M}(x \mid 1, 0)} \propto \frac{p_{E|X,M}(1 \mid x, 1)}{p_{E|X,M}(1 \mid x, 0)} \frac{p_{X|M}(x \mid 1)}{p_{X|M}(x \mid 0)} \tag{3.5}$$

as functions of x.

We now make the assumption that $p_{E|X,M}(1 \mid x, m)$, the probability of enrollment conditioned on the predictor variables and market, is actually independent of the market m once x is fixed. In other words, E and M are conditionally independent given X and $p_{E|X,M}(1 \mid x, m) = p_{E|X}(1 \mid x)$. This seems to be a reasonable assumption since enrollment can be expected to depend on demographic variables such as age, sex, etc., but not to depend on which market the individual belongs to once those demographic variables are specified. With this assumption of conditional independence, (3.5) simplifies to

$$p_{X|E,M}(x \mid 1, 1) \propto p_{X|E,M}(x \mid 1, 0) \frac{p_{X|M}(x \mid 1)}{p_{X|M}(x \mid 0)}.$$

Since the training samples are distributed according to $p_{X|E,M}(x \mid 1, 0)$ while the test samples are distributed according to $p_{X|E,M}(x \mid 1, 1)$, the importance weighting is therefore $p_{X|M}(x \mid 1)/p_{X|M}(x \mid 0)$ (up to a constant of proportionality), taking the place of $q_X(x)/p_X(x)$ in Section 2. Obtaining the required PMFs via empirical distributions in the same way as discussed in Section 2, the weighted empirical risk from (2.3) formally becomes

$$\sum_x \hat{p}_{X|E,M}(x \mid 1, 0) \frac{\hat{p}_{X|M}(x \mid 1)}{\hat{p}_{X|M}(x \mid 0)} \frac{1}{n(x)} \sum_{i: x_i = x} \mathcal{L}(f(x), y_i).$$

3.1 Importance Weighting via Logistic Regression Logistic regression provides a parametric alternative to estimating the probability ratio $q_X(x)/p_X(x)$, i.e. $p_{X|M}(x \mid 1)/p_{X|M}(x \mid 0)$. First, a logistic regression model is trained to decide between the existing market $M = 0$ and new market $M = 1$ given the covariate x, using demographic data for the two markets. The model yields a parametric form for the conditional probability of belonging to each market,

$$p_{M|X}(1 \mid x) = \frac{1}{1 + e^{-\beta^T x}}, \quad p_{M|X}(0 \mid x) = \frac{e^{-\beta^T x}}{1 + e^{-\beta^T x}}.$$

An application of Bayes' rule shows that

$$e^{\beta^T x} = \frac{p_{M|X}(1 \mid x)}{p_{M|X}(0 \mid x)} \propto \frac{p_{X|M}(x \mid 1)}{p_{X|M}(x \mid 0)}$$

as functions of x. Hence the desired probability ratio is given by $e^{\beta^T x}$, the exponential of a linear function of x. It follows that the ratio is only allowed to vary monotonically in the direction β. The advantage of logistic regression is that there are only as many parameters to estimate in β as there are variables in x. The disadvantage is that the estimated probability ratio is highly restricted and may differ significantly from the true ratio.

3.2 Aggregate Prediction Error In health care insurance applications such as market risk assessment, one is typically interested in aggregate predictions of the average or total cost for groups of individuals. For example, one may wish to predict the average cost for a new enrollee population as a whole or for segments of the population. This section overviews calculations relating the aggregate prediction error to common measures of performance.

Suppose that the average cost for m i.i.d. individuals, $(1/m) \sum_{i=1}^m Y_i$, is predicted by averaging the predictions $\hat{Y}(X_i)$ for each individual. The aggregate prediction error is therefore

$$\varepsilon = \frac{1}{m} \sum_{i=1}^m \left(\hat{Y}(X_i) - Y_i \right).$$

The mean error is given by

$$\mathbb{E}[\varepsilon] = \mathbb{E}[\hat{Y}(X)] - \mathbb{E}[Y] = b(\hat{Y}),$$

where $b(\hat{Y})$ denotes the bias of the predictor. Using the definition of the (population) coefficient of determination R^2,

$$R^2 = 1 - \frac{\mathbb{E}\left[\left(\hat{Y}(X) - Y \right)^2 \right]}{\mathrm{var}(Y)},$$

the error variance can be written as

$$\mathrm{var}(\varepsilon) = \frac{1}{m}\left((1-R^2)\,\mathrm{var}(Y) - b(\hat{Y})^2\right).$$

We compare the error ε to the true average cost $(1/m)\sum_{i=1}^{m} Y_i$, which has mean $\mathbb{E}[Y]$ and variance $\mathrm{var}(Y)/m$. It follows that when m is large, the relative error $\bar{\varepsilon}$ is given to first order by $\varepsilon/\mathbb{E}[Y]$, and

$$\mathbb{E}[\bar{\varepsilon}] \approx \bar{b}(\hat{Y}),$$
$$\mathrm{var}(\bar{\varepsilon}) \approx \frac{1}{m}\left((1-R^2)\frac{\mathrm{var}(Y)}{\mathbb{E}[Y]^2} - \bar{b}(\hat{Y})^2\right),$$

where $\bar{b}(\hat{Y})$ is the relative bias.

Note that as m increases, the variance decreases as $1/m$ while the mean remains constant. Thus the bias of the predictor becomes increasingly important as the size of the aggregate group increases. As an illustration, the mean squared relative error is given by

$$\mathbb{E}[\bar{\varepsilon}^2] \approx \frac{m-1}{m}\bar{b}(\hat{Y})^2 + \frac{1-R^2}{m}\frac{\mathrm{var}(Y)}{\mathbb{E}[Y]^2}. \tag{3.6}$$

The ratio $\mathrm{var}(Y)/\mathbb{E}[Y]^2$ is the squared coefficient of variation of Y, with typical values for health care cost ranging from 5 to 10. Thus even for R^2 close to zero, the two terms in the mean squared error are comparable if the relative bias $\bar{b}(\hat{Y})$ is a few percent and $m \sim 10^4$, and the bias dominates if m is larger.

4 Privacy-Preservation for the Market Risk Estimation Workload.

As discussed in the introduction, the privacy of individuals must be protected when working with their personal health cost data. In particular, taking k-anonymity as the notion of privacy, the quasi-identifiers of the original data x must be converted to some other values $\bar{x}$ in a way that the data for an individual cannot be distinguished from at least $k-1$ others. (Note that for simplicity, when we refer to X, Y in this section, we are really referring to $X, Y \mid e = 1, m = 0$, the data available to the insurance company from its members.) With our goal of achieving small aggregate prediction error of cost Y, we not only want the samples $\bar{x}_i$, $i = 1, \ldots, n$ to have the k-anonymity property, but for the regression model learned from $(\bar{x}_i, y_i)$, $i = 1, \ldots, n$ to have small relative bias, large R^2, and good performance in other measures of generalization. With these dual goals in mind, we propose a novel combination of operations inspired by k-member clustering [7] and distribution-preserving quantization with dithering and transformation [13].

The several operations we propose that map (x_i, y_i) to $(\bar{x}_i, y_i)$ are summarized in Figure 1. The original data

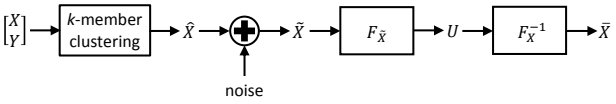


Figure 1: Block diagram of the operations to achieve k-anonymity and distribution preservation.

is first clustered using a modification of the k-member clustering algorithm proposed in [7]. Our modifications are to use Euclidean distance as the distortion criterion and to base the distortion calculation on both the quasi-identifiers X and the sensitive data Y so that individuals with similar costs are grouped together, which is beneficial for the downstream regression. We drop Y once final cluster assignments have been determined. This step achieves our goal of k-anonymity; all other operations that follow are meant to improve the regression model learning. The output of the clustering is the set of values $\hat{x}_i$, where all samples within the same cluster share a $\hat{x}$ value. Let $c = \lfloor \frac{n}{k} \rfloor$ be the number of clusters and j index the clusters. Let $n_j \geq k$ be the number of samples in cluster j.

The output set of the k-member clustering contains only c distinct values that are not distributed like X. The second proposed operation, dithering (the intentional application of noise), returns the data set to having n distinct values. In particular, we estimate covariances of each of the clusters, Σ_j, $j = 1, \ldots, c$, and add Gaussian noise $\mathcal{N}\left(0, \Sigma_{\{j:i\in\mathrm{cluster}_j\}} + \alpha I\right)$ to each sample according to its cluster membership to produce values $\tilde{x}_i$. The extra bit of covariance with parameter $\alpha > 0$ is to account for clusters in which all x values are the same, which tends to occur for smaller values of k. The cumulative distribution function (CDF) of $\tilde{X}$ is a Gaussian mixture with c mixture components:

$$F_{\tilde{X}}(\tilde{x}) = \sum_{j=1}^{c} \frac{n_j}{n}\Phi(\tilde{x}; \hat{x}_j, \Sigma_j + \alpha I). \tag{4.7}$$

This $F_{\tilde{X}}(\cdot)$ may be quite different from the original data distribution $F_{X|E,M}(\cdot \mid 1, 0)$. The goal of the final two operations is to transform $\tilde{X}$ so that it is distributed like X. The vector of quasi-identifiers is in general d-dimensional, $d > 1$, and in our case composed of discrete-valued elements. Due to the multivariate nature of our desired transformation, we require a procedure like the one developed by Rosenblatt [21]. We first use the CDF of $\tilde{X}$ to transform $\tilde{X}$ to a uniformly-distributed variable U and then the inverse CDF of X to transform U to $\bar{X}$, which is distributed like X. Denoting

the lth dimension of a vector with the subscript l,

$$
\begin{aligned}
U_1 &= F_{\tilde{X}_1}(\tilde{X}_1) \\
U_2 &= F_{\tilde{X}_2|\tilde{X}_1}(\tilde{X}_2 \mid \tilde{X}_1) \\
&\vdots \\
U_d &= F_{\tilde{X}_d|\tilde{X}_1,\ldots,\tilde{X}_{d-1}}(\tilde{X}_d \mid \tilde{X}_1, \ldots, \tilde{X}_{d-1}).
\end{aligned} \tag{4.8}
$$

The conditional CDFs are univariate Gaussian mixtures; their parameters and mixture weights can be obtained in closed form from (4.7). The second of the final operations is similar, but with the inverse CDF of X:

$$
\begin{aligned}
\bar{X}_1 &= F^{-1}_{X_1}(U_1) \\
\bar{X}_2 &= F^{-1}_{X_2|X_1}(U_2 \mid U_1) \\
&\vdots \\
\bar{X}_d &= F^{-1}_{X_d|X_1,\ldots,X_{d-1}}(U_d \mid U_1, \ldots, U_{d-1}).
\end{aligned} \tag{4.9}
$$

In practice, the inverse CDFs are empirical estimates and thus, since conditioning reduces the number of samples available on which to base the estimate, it is good to order the dimensions in the sequential procedure in increasing number of discrete values.

At the end of the process, the sensitive y_i values are rejoined with the clustered and transformed quasi-identifiers $\bar{x}_i$. Overall, this sequence of steps yields output samples $(\bar{x}_i, y_i)$ that are as close as possible to the samples (x_i, y_i) in distribution while being k-anonymous. The main free parameter, k, can be varied to achieve the desired tradeoff between privacy and aggregate prediction error.

5 Empirical Results.

In this section, we describe the results of an empirical study on real-world health cost data. Section 5.1 discusses data sources and simulation of the different populations. Sections 5.2 and 5.3 present prediction results without and with privacy constraints, respectively.

5.1 Description of Data We use publicly-available MEPS data, which shares many characteristics with actual health cost data from insurance companies that we have worked with in the recent past, but cannot include in this paper due to its confidentiality. Based on large-scale surveys, MEPS contains the annual health care cost and demographic information of people across the United States. However, since it does not come from an insurance company, there is neither a concept of a market in the data nor of enrollment in a company's plan. Thus in order to perform market risk assessment, we define two market populations and enrolled subsets of these populations as described below.

We consider a scenario in which an insurance company is currently active in many areas that collectively are representative of the US as a whole. The company is deciding whether to enter specific rating areas in California, where a rating area consists of one or more counties. Therefore the demographic distribution of the existing market, $p_{X|M}(x \mid 0)$, can be taken to be that of the US, while the new market distributions $p_{X|M}(x \mid 1)$ correspond to California rating areas. To simulate these two markets, the MEPS dataset is randomly and evenly split into training and test sets. All results reported in Sections 5.2 and 5.3 are averaged over 20 such splits. The existing market distribution $p_{X|M}(x \mid 0)$ is estimated empirically directly from the training set. The distribution $p_{X|M}(x \mid 1)$ is obtained by reweighing samples from the test set according to the demographics of each rating area, relative to the national baseline represented by MEPS. Rating area-specific demographics are obtained from the American Community Survey (ACS) [1].

Once the market distributions are created, the enrollment in the company's plan must also be simulated. We focus on the dependence of enrollment on age. To generate the existing plan distribution $p_{X|E,M}(x \mid 1, 0)$, samples in the existing market dataset are reweighed based on the age distribution in the initial enrollment period of the Health Insurance Marketplaces created by the Affordable Care Act [2, 3], again relative to the national baseline. The resulting distribution differs notably from that of the larger market, $p_{X|M}(x \mid 0)$, in having few children (< 18) and seniors (> 65). The age-dependent enrollment probabilities $p_{E|X}(1 \mid x)$ induced by this procedure are then applied to samples in the new market to simulate plan enrollment in the new market, $p_{X|E,M}(x \mid 1, 1)$.

The specific MEPS dataset we consider is for the year 2005, containing just over 15000 weighted records, and the demographic variables are gender, age (binned into 8 groups similar to those in [3]), education level (0–5), and income level (categories 0–4 relative to the federal poverty level). The specific cost variable we use is known in MEPS as "total expenditure" (TOTEXP) over the year.

5.2 Estimation Results Without Privacy Preservation We first discuss cost prediction in the absence of privacy-preserving data transformations. Two covariate shift methods are compared, the nonparametric method described in Section 3 and the logistic regression method of Section 3.1, along with a baseline that does not account for covariate shift. The predictive model (corresponding to the function class $\mathcal{F}$ in Section 2) used in all cases is a sum of univariate

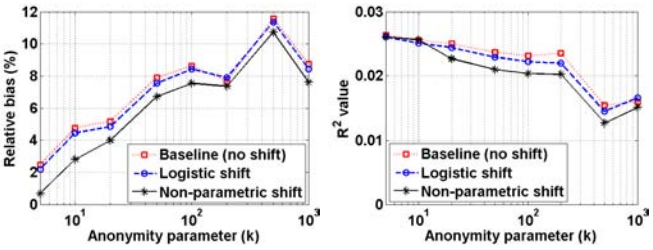
functions of each demographic variable. Note that while the functions are combined linearly, the functions themselves are not constrained to be linear and can vary arbitrarily with input value.

Table 1 summarizes performance in predicting the cost in even-numbered rating areas as new markets. Results for odd-numbered rating areas are similar. As argued in Section 3.2, for aggregate prediction the bias is often the more important performance metric. Table 1 shows that the baseline approach has a noticeable bias, whereas the proposed approaches reduce the bias by shifting the distribution of existing plan members to look more like prospective enrollees in the new market. This reduction is particularly significant with the non-parametric shift method.

5.3 Results With Privacy Preservation Next we present results where the insurer's existing plan data, i.e. $X, Y \mid E = 1, M = 0$, is first subject to privacy-preserving transformation instead of being used directly. We focus in this subsection on rating area 18 as the new market. Two procedures are investigated: the full procedure described in Section 4 in which k-member clustering is followed by a probability transformation to recover the original distribution; and conventional k-anonymization that only includes the clustering and replaces samples of X with corresponding cluster centers $\hat{x}_j$. Privacy transformation may be necessary for the insurer's plan data because healthcare cost is considered potentially sensitive information. However, data for the markets-at-large, both existing and new, does not include cost and hence does not require the same protection.

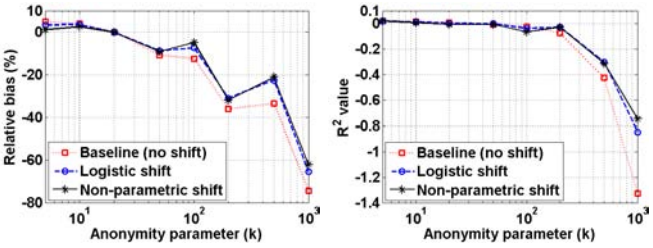
Figure 2 shows the prediction performance of the proposed privacy-preserving procedure combined with different covariant shift methods. In general, as the anonymity parameter k increases, the relationship between X and Y inevitably becomes distorted relative to the original relationship, causing the prediction bias (as defined in Section 3.2) to increase and R^2 to decrease. The non-parametric covariate shift method succeeds at reducing bias, more significantly for small k where the distortion is mild and less so for large k where the distortion can be severe. On the other hand, R^2 is slightly lower for the covariate shift methods because the reweighting of training samples to reduce bias also introduces some additional variability. Note that R^2 is in any case quite low for this difficult healthcare cost prediction problem based on demographics alone, and that for aggregate prediction, R^2 is less important than bias.

Figure 3 shows results for k-anonymization without distribution preservation. As k increases, the original samples from the plan data are represented more and more coarsely by their cluster centers. As a consequence, prediction accuracy suffers markedly.



(a) (b)

Figure 2: Prediction bias (a) and R^2 coefficient (b) resulting from different covariate shift methods and the proposed distribution-preserving procedure to achieve k-anonymity for different values of k. As k increases, distribution preservation moderates the increase in bias, the more important metric for aggregate prediction, while the proposed three-population covariate shift methods reduce bias.



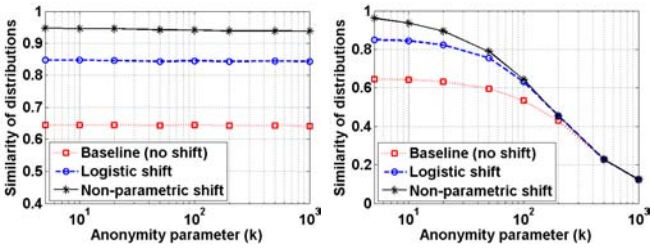
(a) (b)

Figure 3: Bias (a) and R^2 coefficient (b) for different covariate shift methods and k-anonymization without distribution preservation. Prediction error increases unacceptably as k increases.

Some insight into the behavior in Figures 2 and 3 can be seen in Figure 4, which depicts the similarity between the new enrollment distribution estimated through the covariate shift methods, and the actual new enrollment (as simulated). We use the histogram intersection similarity [24] for concreteness. A value close to 1 implies that the predictor is trained on a distribution much like the one encountered in testing. Using distribution-preserving privacy transformations in Figure 4(a), the similarity can be kept constant as the anonymity k increases and can be further enhanced by the covariate shift methods. However, under conventional k-anonymization in Figure 4(b) the similarity decreases rapidly with k.

Table 1: Coefficient of determination R^2 and relative bias when new markets are chosen as rating areas of California, for the baseline (no shift) method and the proposed logistic and non-parametric shift methods.

New Market	R^2 value			Relative Bias (%)		
	No Shift	Logistic	Non-param.	No Shift	Logistic	Non-param.
RA 2	0.0247	0.0225	0.0226	7.53	6.86	3.60
RA 4	0.0270	0.0252	0.0252	4.42	3.90	2.63
RA 6	0.0262	0.0243	0.0244	4.23	3.41	2.21
RA 8	0.0251	0.0235	0.0233	4.14	3.18	1.93
RA 10	0.0257	0.0242	0.0241	4.41	3.56	1.95
RA 12	0.0259	0.0242	0.0243	4.57	3.74	1.91
RA 14	0.0271	0.0253	0.0245	4.43	3.71	2.20
RA 15-16	0.0291	0.0275	0.0283	2.53	2.28	0.32
RA 18	0.0247	0.0245	0.0245	3.04	2.04	0.44



(a) (b)

Figure 4: Histogram intersection between the estimated and actual new enrollment distributions for different covariate shift methods. In (a), the proposed distribution-preserving transformation maintains constant similarities as the anonymity k increases, while in (b), the similarity deteriorates under conventional k-anonymization.

6 Conclusion.

In this paper we have addressed the market risk assessment problem, especially pertinent after passage of the Affordable Care Act, in which insurance companies decide which markets to enter based on the expected cost distribution of enrollees in the new market. To date, insurance companies have not developed sophisticated data mining and machine learning approaches for this problem, only relying on crude estimates mainly driven by intuition and very coarse-level aggregate data. The main issue is the lack of health cost data from markets in which insurers are not active. To solve this problem and allow the use of more advanced regression methods and individual member-level data, we develop two versions of a novel three-population covariate shift-based estimation procedure. The non-parametric version in particular is successful in significantly reducing the relative bias of the regression that would occur if the covariate shift were not done, as seen using real-world MEPS data with realistic simulations for new markets and enrollment probabilities.

Using the accurate proposed approach with fine-level member health cost data presents legal difficulties for insurance companies due to privacy regulations in the United States. Therefore, to construct a full data mining solution that could be deployed, we also propose a novel privacy-preservation method based on dithered quantization and Rosenblatt's transformation. The novelty is required because the workload for the data set after privacy transformation, namely covariate shift and regression, requires preservation of the joint distribution. To the best of our knowledge, such a workload has not been dealt with in the privacy-preserving data mining literature. Our proposed approach is able to maintain predictive accuracy and bias reduction in the downstream regression significantly better than existing privacy-preservation with a non-specific workload. In fact, without our new privacy preservation method, cost prediction essentially fails for k-anonymity greater than ten or twenty.

One direction for future work addresses the following issue. In the market risk assessment problem, it is possible that the conditional distribution $p_{Y|X}$ is not the same in the training and test populations, i.e., current and new markets, unlike in the standard covariate shift problem. For example, the overall cost of living in the new market may differ from that in the existing market and this may affect health care costs as well. However, it is unlikely for there to be sufficient data to learn the full conditional distribution $p_{Y|X,M}$ (otherwise market shift would not be much of a problem). One approximation is to assume a simple scaling where an underlying conditional distribution $p_{Y|X}$ is scaled by a cost-of-living factor $a(M)$ depending on M (implying that the conditional mean for example is $\mathbb{E}[Y \mid X, M] = a(M)\mathbb{E}[Y \mid X]$).

Another direction for future work is theoretical analysis of the proposed privacy-preservation method. Dithered quantization has much supporting theory that

we would like to further explore in the context of privacy, where it has never been applied before. Furthermore, although this paper is wholly concerned with the healthcare domain, similar problems occur in other applications where privacy is regulated. For example in education, charter schools need to estimate properties of markets they may enter, and data is protected by the Family Educational Rights and Privacy Act. We would also like to explore stronger notions of privacy such as l-diversity [16] and t-closeness [14] with distribution-preserving workloads.

Acknowledgment

The authors thank A. Gkoulalas-Divanis, V. S. Iyengar, A. Mojsilović, and G. Yuen-Reed for conversations and support.

References

[1] *American Community Survey*, 2005. United States Census Bureau.

[2] *Enrollment in the health insurance marketplace totals over 8 million people*, press release, United States Department of Health and Human Services, May 2014.

[3] *Health insurance marketplace: Summary enrollment report for the initial annual open enrollment*, ASPE issue brief, United States Department of Health and Human Services, May 2014.

[4] J. Aitchison, *On the distribution of a positive random variable having a discrete probability mass at the origin*, J. Am. Stat. Assoc., 50 (1955), pp. 901–908.

[5] M. Alamgir, G. Lugosi, and U. von Luxburg, *Density-preserving quantization with application to graph downsampling*, in Proc. Conf. Learn. Theory, Barcelona, Spain, June 2014, pp. 543–559.

[6] A. Basu and W. G. Manning, *Issues for the next generation of health care cost analyses*, Med. Care, 47 (2009), pp. S109–S114.

[7] J.-W. Byun, A. Kamra, E. Bertino, and N. Li, *Efficient k-anonymization using clustering techniques*, in Proc. Int. Conf. Database Syst. Adv. Appl., Bangkok, Thailand, Apr. 2007, pp. 188–200.

[8] P. Diehr, D. Yanez, A. Ash, M. Hornbrook, and D. Y. Lin, *Methods for analyzing health care utilization and costs*, Annu. Rev. Public Health, 20 (1999), pp. 125–144.

[9] R. M. Gray and D. L. Neuhoff, *Quantization*, IEEE Trans. Inf. Theory, 44 (1998), pp. 2325–2383.

[10] V. S. Iyengar, *Transforming data to satisfy privacy constraints*, in Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min., Edmonton, Canada, July 2002, pp. 279–288.

[11] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, *On the privacy preserving properties of random data perturbation techniques*, in Proc. IEEE Int. Conf. Data Min., Melbourne, FL, Nov. 2003, pp. 99–106.

[12] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, *Mondrian multidimensional k-anonymity*, in Proc. IEEE Int. Conf. Data Eng., Atlanta, GA, Apr. 2006.

[13] M. Li, J. Klejsa, and W. B. Kleijn, *Distribution preserving quantization with dithering and transformation*, IEEE Signal Process. Lett., 17 (2010), pp. 1014–1017.

[14] N. Li, T. Li, and S. Venkatasubramanian, *t-closeness: Privacy beyond k-anonymity and l-diversity.*, in Proc. IEEE Int. Conf. Data Eng., Istanbul, Turkey, Apr. 2007, pp. 106–115.

[15] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, *Quantization and dither: A theoretical survey*, J. Audio Eng. Soc., 40 (1992), pp. 355–375.

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, *l-diversity: Privacy beyond k-anonymity*, ACM Trans. Knowl. Disc. Data, 1 (2007), p. 3.

[17] B. Malin, K. Benitez, and D. Masys, *Never too old for anonymity: A statistical standard for demographic data sharing via the HIPAA privacy rule*, J. Am. Med. Inform. Assoc., 18 (2011), pp. 3–10.

[18] D. G. Messerschmitt, *Quantizing for maximum output entropy*, IEEE Trans. Inf. Theory, IT-17 (1971), p. 612.

[19] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, eds., *Dataset Shift in Machine Learning*, MIT Press, Cambridge, MA, 2009.

[20] D. Rebollo-Monedero, J. Forné, E. Pallarès, and J. Parra-Arnau, *A modification of the Lloyd algorithm for k-anonymous quantization*, Inf. Sci., 222 (2013), pp. 185–202.

[21] M. Rosenblatt, *Remarks on a multivariate transformation*, Ann. Math. Stat., 23 (1952), pp. 470–472.

[22] P. Samarati, *Protecting respondents' identities in microdata release*, IEEE Trans. Knowl. Data Eng., 13 (2001), pp. 1010–1027.

[23] M. Sugiyama, M. Krauledat, and K.-R. Müller, *Covariate shift adaptation by importance weighted cross validation*, J. Mach. Learn. Res., 8 (2007), pp. 985–1005.

[24] M. J. Swain and D. H. Ballard, *Color indexing*, Int. J. Comput. Vis., 7 (1991), pp. 11–32.

[25] L. Sweeney, *k-anonymity: A model for protecting privacy*, Int. J. Uncertain. Fuzz., 10 (2002), pp. 557–570.

[26] D. Wei, K. N. Ramamurthy, D. A. Katz-Rogozhnikov, and A. Mojsilović, *Multiplicative regression via constrained least squares*, in Proc. IEEE Workshop Stat. Signal Process., Gold Coast, Australia, Jun.–Jul. 2014, pp. 304–307.

[27] J. Yi, J. Wang, and R. Jin, *Privacy and regression model preserved learning*, in Proc. AAAI Conf. Artif. Intell., Québec City, Canada, July 2014, pp. 1341–1347.