Received: 10 September 2015 | Revised: 27 October 2016 | Accepted: 1 February 2018

DOI: 10.1002/sam.11374

ORIGINAL ARTICLE

Distribution-preserving *k*-anonymity

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

IBM Research AI, Thomas J. Watson Research Center, Yorktown Heights, New York

Correspondence
Kush R. Varshney, IBM Research AI, Thomas J. Watson Research Center, Yorktown Heights, NY.
Email: krvarshn@us.ibm.com

Preserving the privacy of individuals by protecting their sensitive attributes is an important consideration during microdata release. However, it is equally important to preserve the quality or utility of the data for at least some targeted workloads. We propose a novel framework for privacy preservation based on the *k*-anonymity model that is ideally suited for workloads that require preserving the probability distribution of the quasi-identifier variables in the data. Our framework combines the principles of distribution-preserving quantization and *k*-member clustering, and we specialize it to 2 variants that respectively use intra-cluster and Gaussian dithering of cluster centers to achieve distribution preservation. We perform theoretical analysis of the proposed schemes in terms of distribution preservation, and describe their utility in workloads such as covariate shift and transfer learning where such a property is necessary. Using extensive experiments on real-world Medical Expenditure Panel Survey data, we demonstrate the merits of our algorithms over standard *k*-anonymization for a hallmark health care application where an insurance company wishes to understand the risk in entering a new market. Furthermore, by empirically quantifying the reidentification risk, we also show that the proposed approaches indeed maintain *k*-anonymity.

KEYWORDS

covariate shift, distribution preservation, dithering, health care, *k*-member clustering, microdata release, privacy, reidentifcation risk, Rosenblatt's transformation, supervised learning, transfer learning

1 | INTRODUCTION

Data owners often employ data analysts to provide them with accurate and actionable insights. In many important domains, the data to be analyzed consists of individual records with sensitive fields that must be protected to maintain privacy. The passing of data from owner to analyst, known as microdata release or publishing, necessitates anonymization or some other similar privacy protection operation in these applications. A commonly used privacy criterion in microdata publishing is *k*-anonymity [31,35]. Any operation on a data set whose result achieves *k*-anonymity is equally good from the privacy perspective; it is the *workload* for which the data is to be used that defines the quality or utility of the operation [18,31,39].

A key tool for the data analyst in developing insights is supervised learning. In certain challenging settings, labeled training data is available to the supervised learning algorithm for one problem or set of conditions, but testing is to be performed on a different problem or set of conditions. One common way of addressing this challenge is by *transfer learning* and *covariate shift* methods that reweight training samples in accordance with the probability distributions of the different problems [8,27,33]. In this paper, our contribution is to develop the first (to the best of our knowledge) *k*-anonymity approach for workloads such as transfer learning and covariate shift that require the preservation of the probability distributions of the data [38].

Portions of this work were first presented at the 2015 SIAM International Conference on Data Mining.

Consider a number of colleges with different student distributions that want to release their data to enable the accurate learning of predictive models for a common learning outcome [4]. Using the data from all of the colleges together allows for better learning than separately due to the advantages of transfer learning, but releasing such information is fraught with privacy issues such as satisfying the Family Educational Rights and Privacy Act (FERPA) in the United States [10]. There are similar laws in applications besides education, including the Health Insurance Portability and Accountability Act (HIPAA) in health care and the Gramm-Leach-Bliley Act (GLBA) in personal finance. Several statistical interpretations of the legal language for protecting privacy exist; the property of *k*-anonymity is a common interpretation to lower the risk of reidentification [13,25].

As another example, consider the change in the landscape of health insurance in the United States after the passage of the Patient Protection and Affordable Care Act. 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, that is, low-cost, individuals likely to enroll in their plans. This objective gives rise to the problem of market risk assessment, the estimation of cost profiles of new market populations. Successful market risk assessment, however, faces at least 2 challenges. On the privacy front, even internal use of individual health data by an insurance company for its planning and strategy is protected by HIPAA. On the supervised learning front, the companies had access to member health cost data from their existing markets available for training, but no such cost data available for the new markets, necessitating predictive models with covariate shift.

In privacy, there are 3 types of variables: key attributes, quasi-identifiers, and sensitive attributes. Key attributes, such as the name of the person, are always dropped before microdata release. Quasi-identifiers are often demographic variables like age, gender, and postal code that can be matched to other publicly available data sets such as voter registration records containing both key attributes and quasi-identifiers to reveal identities in the data set to be protected. It is this type of identity disclosure that *k*-anonymity aims to protect against. Under the *k*-anonymity privacy model, the quasi-identifiers for an individual cannot be distinguished from the quasi-identifiers of at least $k - 1$ other individuals [31,35]. The sensitive attributes are variables such as educational or medical test results.

There are many anonymization algorithms to transform the quasi-identifier attributes of a data set to achieve *k*-anonymity, but these existing algorithms do not seek to preserve the data distribution, as is required for covariate shift and transfer learning. Existing algorithms for *k*-anonymity include generalizations and suppressions [18], multidimensional generalization [20], and multidimensional clustering in which the samples or records in the data are grouped by similarity such that the smallest group has at least *k* elements [9]. Clustering approaches offer the most flexibility and best performance.

The reason that existing clustering approaches to anonymization are not distribution-preserving is as follows: the optimization criterion is based on the average distortion of the individual samples and with such a criterion, the optimal cluster centers do not follow the same distribution as the original data. In the asymptotic limit as the number of clusters goes to infinity, the distribution of the cluster center locations is the original data distribution to the one-third power, properly normalized [17].

Distribution-preserving quantization is an alternative to standard clustering methods that does have the desired output behavior [1,22]. It has been developed in the context of audio signal processing and has never been considered in the privacy preservation context before. Specifically, the approach of [22] is based on subtractive dithered quantization [23] followed by Rosenblatt's transformation [29]. We emphasize that although dithering, the introduction of noise or random perturbations, is fraught with several issues when used for privacy preservation [19], in our work, the introduction of noise is not for anonymization purposes, but to allow the manipulation of the distribution. Since distribution-preserving quantization comes from the signal processing and communications domain, not the privacy domain, it does not aim to achieve *k*-anonymity. In particular, like *k*-means clustering and other standard clustering methods, it takes 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.

Clustering for *k*-anonymity requires a different problem setup, given the name *k*-member clustering in Ref. [9]. The *k*-member clustering problem has the *k* of *k*-anonymity as the input parameter rather than the number of clusters. Algorithms for *k*-member clustering may be divided into two classes: objective-driven optimization algorithms and simple scalable algorithms. The first class includes algorithms based on minimum-cost network flow [11] and a cluster penalty function [28]. The second class includes greedy clustering [9], subsampling and local optimization [5], and constrained agglomerative clustering [15]. This problem is also related to clustering with maximum cluster size constraints [14,16], semi-supervised clustering [7], and maximum output entropy quantization [26].

1.1 | Contribution

To the best of our knowledge, there is no existing approach that transforms quasi-identifier data to achieve *k*-anonymity while also preserving its probability distribution. Our solution combines the key aspects of *k*-member clustering algorithms with the key aspects of distribution-preserving quantization to obtain a distribution-preserving *k*-anonymity transformation.

We propose 2 ways to dither cluster centers: intra-cluster dither and Gaussian dither. For each, we theoretically analyze the distribution-preserving properties. As the distribution preservation is motivated by workloads such as covariate shift and transfer learning that require the data distribution, we detail how our proposed method can be used in conjunction with those machine learning tasks. We demonstrate the proposed approach on real-world Medical Expenditure Panel Survey (MEPS) data for the health insurance market risk assessment application mentioned above. We find that the proposed method does in fact maintain k-anonymity empirically by investigating reidentification risk and find that the distribution preservation is critical to obtaining usable machine learning predictions.

The remainder of the paper is organized as follows. In Section 2, we provide background on the transfer learning and covariate shift problems encountered in supervised learning. In Section 3, we develop the new privacy-preservation method for the workloads of interest that combines aspects of k-member clustering and distribution-preserving quantization. In Section 4, we provide empirical results on real-world health care data for market risk assessment. Section 5 provides a summary and discussion.

2 | COVARIATE SHIFT AND TRANSFER LEARNING

In this section, we first introduce notation, then describe the basic learning problem encountered in the covariate shift setting, and finally describe the problem in the transfer learning setting. Random variables are indicated by capital letters and their samples by lowercase letters. $\mathbb{E}$ is used to denote the expectation of a random variable.

Consider the following general supervised learning 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,

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

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. In the most idealized setting, covariate shift does not necessarily pose a problem: As the number of samples $n \to \infty$, the empirical risk in Equation (1) converges to the population risk

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

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, in practical settings where n is finite or $\mathcal{F}$ is overly constrained, then the predictor $\widehat{Y}$ resulting from Equation (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

$$w(x) = \frac{q_X(x)}{p_X(x)}.$$

This weighting represents the relative importance of each sample under q_X rather than p_X. The weighted empirical risk

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

then converges to

$$\mathbb{E}_{p_X p_{Y|X}}[w(X)\mathcal{L}(f(X), Y)] = \mathbb{E}_{q_X p_{Y|X}}[\mathcal{L}(f(X), Y)], \tag{2}$$

thus matching the test distribution.

In the transfer learning problem, we have m different tasks each with their own joint distributions in training and testing: $p_{X,Y|T}$ and $q_{X,Y|T}$, where $T \in \{1, \ldots, m\}$ is a variable indicating the task. The tasks may be related to each other and we would like to use the training samples from all other tasks in helping learn f_t, the predictor for task t. Like covariate shift, the transfer learning problem may be addressed by distribution matching through the weight

$$w(x, y \mid T = t) = \frac{p_{X,Y|T}(x, y \mid T = t)}{\sum\limits_{t'=1}^{m} p_T(T = t') p_{X,Y|T}(x, y \mid T = t')} \times \frac{q_{X|T}(x \mid T = t)}{p_{X|T}(x \mid T = t)}.$$

Note that the second part of the weight is the same as the covariate shift weight to take differences between training and testing into account. The first part of the weight allows for the transfer of information across tasks. With this weight, we have convergence to [8]:

$$\mathbb{E}_{\sum_{t'=1}^{m} p_{t'} p_{X|t'} p_{Y|X,t'}}[w(X, Y \mid t)\mathcal{L}(f_t(X), Y)] = \mathbb{E}_{q_{X|t} p_{Y|X,t}}[\mathcal{L}(f_t(X), Y)], \tag{3}$$

thereby matching the test distribution of task t.

We must also estimate the weights $w(x)$ and $w(x, y|t)$ that we have defined in the learning process. There are a variety of ways to estimate them including nonparametric methods and methods based on logistic regression, cf. [8,38]. We do not go into the details here, but only emphasize that the data distributions are the key ingredient when estimating the weights. Therefore, any privacy-preservation operation performed on the data should ideally preserve the data distributions.

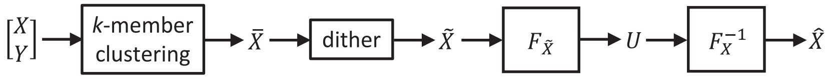

FIGURE 1 Block diagram of the operations to achieve k-anonymity and distribution preservation

3 | DISTRIBUTION-PRESERVING *k*-ANONYMITY

As discussed in the introduction, the privacy of individuals must be protected when working with their personal data in domains such as education and health care. In particular, taking k-anonymity as the notion of privacy, the quasi-identifiers x in the original data must be converted to some other values $\hat{x}$ in a way that the data for an individual cannot be distinguished from at least $k-1$ others. (For notational simplicity, we use x, y generically in this section to refer to training data distributed as $p_{X,\,Y}$ in the standard supervised learning or covariate shift settings, or to each training data set distributed as $p_{X,\,Y|T=t}$, $t = 1, \ldots, m$ in the transfer learning setting.) Moreover, given our ultimate goal of predicting response Y, we not only want the samples $\hat{x}_i$, $i = 1, \ldots, n$ to have the k-anonymity property, but also the model learned from $(\hat{x}_i, y_i)$, $i = 1, \ldots, n$ to have small prediction error, as quantified by relative bias, R^2, or other measures of generalization.

With these dual goals in mind, in Ref. [38] we introduced a sequence of operations inspired by k-member clustering [9] and distribution-preserving quantization with dithering and transformation [22]. In the present paper, we generalize and further develop this framework. Figure 1 illustrates the overall procedure. Section 3.1 discusses the first step of k-member clustering while Section 3.2 discusses the subsequent distribution-recovering transformation.

3.1 | *k*-member clustering

The original data (x_i, y_i), $i = 1, \ldots, n$ is first clustered subject to the k-anonymity requirement that each cluster contain at least k members. Define c to be the number of clusters and let $\ell = 1, \ldots, c$ index the clusters. Let $a_{i\ell} = 1$ if sample i is assigned to cluster ℓ and $a_{i\ell} = 0$ otherwise. Denote by $(\bar{x}_\ell, \bar{y}_\ell)$ the centroid chosen to represent the samples in cluster ℓ. The centroids are determined in part by a function $d((x_i, y_i), (\bar{x}_\ell, \bar{y}_\ell))$ that measures the distortion between 2 points. Here we choose d to be a weighted squared Euclidean distance,

$$d((x_i, y_i), (\bar{x}_\ell, \bar{y}_\ell)) = \|x_i - \bar{x}_\ell\|_2^2 + w(y_i - \bar{y}_\ell)^2, \tag{4}$$

where $w > 0$ is the weight on the response component relative to the quasi-identifier components, intended to account for any scale difference between them.

With the foregoing definitions, k-member clustering can be formulated as the following optimization problem:

$$\begin{aligned} \min_{\{a_{i\ell}\},\{(\bar{x}_\ell,\bar{y}_\ell)\}} \quad & \sum_{i=1}^{n}\sum_{\ell=1}^{c} a_{i\ell}\, d((x_i, y_i), (\bar{x}_\ell, \bar{y}_\ell)) \\ \text{s.t.} \quad & a_{i\ell} \in \{0, 1\} \qquad \forall\ i, \ell, \\ & \sum_{\ell=1}^{c} a_{i\ell} = 1 \qquad \forall\ i, \\ & \sum_{i=1}^{n} a_{i\ell} \geq k \qquad \forall\ \ell. \end{aligned} \tag{5}$$

The last line in Equation (5) expresses the k-member constraint while the second-last line ensures that each sample is assigned to exactly 1 cluster. These constraints imply that the number of clusters c is at most $\left\lfloor \frac{n}{k} \right\rfloor$. In this work, we simply fix $c = \left\lfloor \frac{n}{k} \right\rfloor$; optimization over the number of clusters is a subject for future study.

Solutions to the k-member clustering problem Equation (5) can be obtained using methods such as Refs. [5,9,11,15,28] discussed in Section 1. In the experiments of Section 4, due to its scalability, we use the greedy algorithm proposed in Ref. [9] modified to use the Euclidean distortion function (Equation (4)).

Once a solution to Equation (5) has been determined, the original response values y_i are set aside, to be rejoined only at the end with the transformed quasi-identifiers $\hat{x}_i$. In place of each original x_i, we record only the cluster ℓ_i to which it was assigned, that is, $\ell_i = \ell$ for which $a_{i\ell} = 1$. For each cluster ℓ, the associated set of values $\mathscr{V}_\ell = \{x_i,\ i : a_{i\ell} = 1\}$ is retained; the centroid $\bar{x}_\ell$ is a function of $\mathscr{V}_\ell$ and may or may not be retained depending on the type of dither to be applied. Importantly however, an index i is no longer linked to any single value in $\mathscr{V}_{\ell_i}$. In this way the clustering step achieves our goal of k-anonymity; since all operations that follow depend only on the whole sets $\mathscr{V}_{\ell_i}$, a level of privacy equivalent to k-anonymity is maintained. In Section 4.3, we confirm through reidentification experiments that the proposed distribution-recovering operations not only preserve k-anonymity but can also result in even stronger privacy.

3.2 | Distribution-recovering transformation

The operations following k-member clustering in Figure 1 aim to reconstruct a k-anonymous data set that follows the distribution of X. Using the clustering output described at the end of Section 3.1, the first operation of dithering (the intentional application of noise) produces $\tilde{x}_i$, $i = 1, \ldots, n$, sampled from a continuous probability distribution. (By contrast, simply using the centroids $\bar{x}_\ell$ would result in a discrete distribution.) Different types of dither give rise to different variants of the proposed procedure. This makes our framework quite general since any dither with a continuous distribution may be used. In Sections 3.2.1 and 3.2.2, we discuss 2 versions corresponding to intra-cluster and Gaussian dither.

The final 2 operations in Figure 1 constitute Rosenblatt's transformation [29] for transforming an arbitrary continuous distribution into a specified target distribution, in our case that of X. The first of these operations, transformation to a uniform distribution, is given by the following sequence of (conditional) cumulative distribution functions (CDF) of $\tilde{X}$, one for each dimension:

$$\begin{aligned} u_{i1} &= F_{\tilde{X}_1}(\tilde{x}_{i1}) = F_{\tilde{X}_1|\tilde{X}^0}(\tilde{x}_{i1} \mid \tilde{x}_i^0) \\ u_{i2} &= F_{\tilde{X}_2|\tilde{X}_1}(\tilde{x}_{i2} \mid \tilde{x}_{i1}) = F_{\tilde{X}_2|\tilde{X}^1}(\tilde{x}_{i2} \mid \tilde{x}_i^1) \\ &\vdots \\ u_{id} &= F_{\tilde{X}_d|\tilde{X}^{d-1}}(\tilde{x}_{id} \mid \tilde{x}_i^{d-1}), \end{aligned} \tag{6}$$

where $\tilde{x}_i^{j-1}$ is shorthand for $\tilde{x}_{i1}, \ldots, \tilde{x}_{i,j-1}$ (similarly for other multidimensional quantities) and a superscript of 0 is understood to indicate the empty list. The detailed forms of the CDFs in Equation (6) depends on the choice of dither and are specified in Sections 3.2.1 and 3.2.2. Transformation Equation (6) is applied to all samples $i = 1, \ldots, n$ and can be done in parallel, for example cluster by cluster.

The last operation, transformation from a uniform distribution to the distribution of X, is similar to Equation (6) but with the inverse CDF of X:

$$\begin{aligned} \hat{x}_{i1} &= F^{-1}_{X_1}(u_{i1}) = F^{-1}_{X_1|X^0}(u_{i1} \mid \hat{x}_i^0) \\ \hat{x}_{i2} &= F^{-1}_{X_2|X_1}(u_{i2} \mid \hat{x}_{i1}) = F^{-1}_{X_2|X^1}(u_{i2} \mid \hat{x}_i^1) \\ &\vdots \\ \hat{x}_{id} &= F^{-1}_{X_d|X^{d-1}}(u_{id} \mid \hat{x}_i^{d-1}), \end{aligned} \tag{7}$$

where the generalized inverse CDF is defined as

$$F_X^{-1}(u) = \inf\{x : F_X(x) \geq u\}. \tag{8}$$

In practice, the underlying distribution that generates the data x is not known. Instead we use the empirical distribution, in which case X can always be regarded as discrete. For each dimension $j = 1, \ldots, d$, define $v_j(1) < v_j(2) < \ldots < v_j(n_j)$, $n_j \leq n$, to be the distinct observed values of X_j in increasing order. Then by specializing Equation (8) to discrete distributions, transformation Equation (7) can be expressed more explicitly as

$$\begin{gathered} \hat{x}_{ij} = v_j(i_j) \qquad \text{if} \\ F_{X_j|X^{j-1}}(v_j(i_j - 1) \mid \hat{x}_i^{j-1}) < u_{ij} \leq F_{X_j|X^{j-1}}(v_j(i_j) \mid \hat{x}_i^{j-1}), \\ i_j = 1, \ldots, n_j, \qquad j = 1, \ldots, d, \end{gathered} \tag{9}$$

where we define $v_j(0) = -\infty$ for $i_j = 1$ so that $F_{X_j|X^{j-1}}(v_j(0) \mid \hat{x}_i^{j-1}) = 0$.

At the end of the process, the sensitive y_i values are rejoined with the clustered and transformed quasi-identifiers $\hat{x}_i$. Overall, this sequence of steps yields output samples $(\hat{x}_i, y_i)$, $i = 1, \ldots, n$ that are close to the original 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 prediction error.

3.2.1 | Intra-cluster dither

In this subsection, we consider dither distributions that are supported only on the cluster to which a sample belongs. In other words, for each i, the cluster assignment ℓ_i is viewed as a conditioning event and $\tilde{x}_i$ is sampled randomly from the support of ℓ_i with probability 1. Here the support can be any subset of $\mathbb{R}^d$ that contains the points in $\mathcal{V}_{\ell_i}$ and no others.

The framework of Figure 1 with intra-cluster dither as defined above is still quite general since it leaves open the exact specification of support sets and corresponding probability distributions. In the following we show that it encompasses the important special case of random resampling with replacement from the values within each cluster. In this approach, each transformed sample $\hat{x}_i$ is drawn randomly from the value set $\mathcal{V}_{\ell_i}$ of the corresponding cluster. A closely related alternative is resampling *without* replacement, in which $\hat{x}_i$ corresponding to the same cluster ℓ_i are chosen as a random permutation of $\mathcal{V}_{\ell_i}$.

We focus here on resampling with replacement under the constraint that $\hat{X}$ preserves the distribution of X, as in Figure 1. This constraint can be met by selecting values within a cluster according to their empirical frequencies, as stated in Lemma 1. First define $v(i_1, \ldots, i_d) = (v_1(i_1), \ldots, v_d(i_d))$ to be the multidimensional extension of $v_j(i_j)$, noting that some of these combinations of values may not be observed. Denote by $n_\ell(i_1, \ldots, i_d)$ the number of samples in cluster ℓ with value $v(i_1, \ldots, i_d)$; $n(i_1, \ldots, i_d)$ (without the subscript ℓ) denotes the corresponding number in all clusters; n_ℓ (without the indices i_j) denotes the number of samples in cluster ℓ ($n_\ell \geq k$ from Equation (5)); and n is the total number of samples as before. Some of these definitions are illustrated in Figure 2.

Lemma 1. If for all samples i and conditioned on cluster ℓ_i, $\hat{x}_i = v(i_1, \ldots, i_d)$ is chosen with probability $n_{\ell_i}(i_1, \ldots, i_d)/n_{\ell_i}$, then the transformed quasi-identifiers $\hat{X}$ have the same distribution as X.

Proof. The lemma follows from calculating the probability mass function (PMF, denoted generically by p) of $\hat{X}$. Treating cluster membership as a conditioning random variable L, we have

$$\begin{aligned} p_{\hat{X}}(v(i_1, \ldots, i_d)) &= \sum_{\ell=1}^{c} p_L(\ell) p_{\hat{X}|L}(v(i_1, \ldots, i_d) \mid \ell) \\ &= \sum_{\ell=1}^{c} \frac{n_\ell}{n} \frac{n_\ell(i_1, \ldots, i_d)}{n_\ell} \\ &= \frac{n(i_1, \ldots, i_d)}{n} \\ &= p_X(v(i_1, \ldots, i_d)), \end{aligned}$$

using the definition of empirical distribution in the last line. □

To establish the equivalence between the proposed procedure and resampling with replacement, we construct a rectangular partition of $\mathbb{R}^d$ and define support sets accordingly.

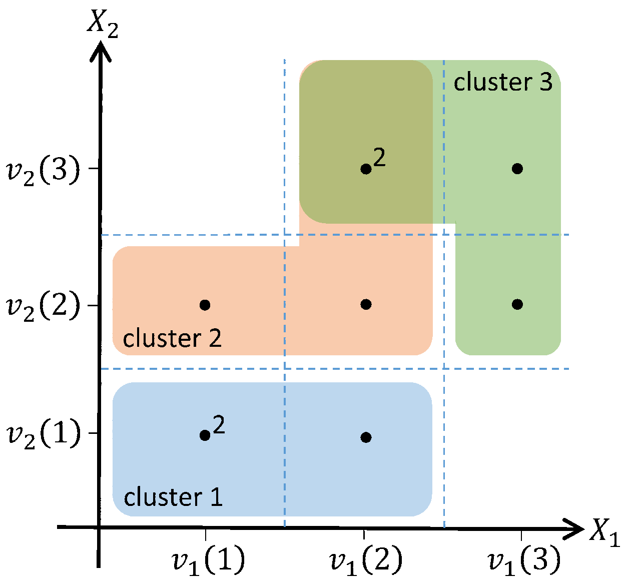

FIGURE 2 Partition of a 2-dimensional ($d = 2$) quasi-identifier space into rectangular cells $\mathscr{C}(i_1, i_2)$, each containing a single value $v(i_1, i_2)$. Different shadings indicate the support sets of different clusters, which may overlap. Cells $\mathscr{C}(1, 1)$ and $\mathscr{C}(2, 3)$ each contain $n(1, 1) = n(2, 3) = 2$ samples; all other cells with dots contain $n(i_1, i_2) = 1$ sample. In (1, 1), both samples belong to cluster 1 ($n_1(1, 1) = 2$), while in (2, 3), the samples are split between clusters 2 and 3 ($n_2(2, 3) = n_3(2, 3) = 1$)

Figure 2 provides an illustration in the 2-dimensional case $d = 2$. For $j = 1, \dots, d$, let $\{\mathcal{I}_j(i_j)\}_{i_j=1}^{n_j}$ be a set of intervals that partition the real line such that each $\mathcal{I}_j(i_j)$ contains the observed value $v_j(i_j)$. These intervals could be defined for example by the midpoints between the $v_j(i_j)$. The full d-dimensional space is partitioned into rectangular cells $\mathscr{C}(i_1, \dots, i_d) = \mathcal{I}_1(i_1) \times \mathcal{I}_2(i_2) \times \dots \times \mathcal{I}_d(i_d)$ formed by Cartesian products of intervals. Each cell $\mathscr{C}(i_1, \dots, i_d)$ contains a single $v(i_1, \dots, i_d)$. The *support* $\mathscr{C}_\ell$ of cluster ℓ is then defined as the union of cells $\mathscr{C}(i_1, \dots, i_d)$ such that the corresponding $v(i_1, \dots, i_d) \in \mathscr{V}_\ell$.

We now define a dither distribution by the probabilities

$$\Pr(\tilde{X} \in \mathscr{C}(i_1, \dots, i_d) \mid \ell) = \frac{n_\ell(i_1, \dots, i_d)}{n_\ell}. \qquad (10)$$

Conditioned on $\tilde{X} \in \mathscr{C}(i_1, \dots, i_d)$ (and independent of cluster ℓ), let $\tilde{X}$ be uniformly distributed over $\mathscr{C}(i_1, \dots, i_d)$.

Theorem 1. For dither $\tilde{X}$ with piecewise-uniform conditional distribution defined by Equation (10), the procedure in Figure 1 is equivalent to resampling with replacement from the cluster values in $\mathscr{V}_\ell$ with probabilities $n_\ell(i_1, \dots, i_d)/n_\ell$.

Proof. Please see Appendix A. □

In the proof of Theorem 2, specifically in going from Equations (A3) to (A4), the assumption of a piecewise-uniform dither distribution is not strictly necessary. We have chosen to make this assumption because it simplifies the proof and interpretation. Furthermore, in the 1-dimensional case $d = 1$, the proof makes no reference to uniformity or any property of the cell-conditional distributions other than continuity. In fact, the cell probability requirements (Equation (10)) can be further relaxed, specifically by merging contiguous cells that all belong in full to the same cluster, that is, cells such that $n(i_1) = n_\ell(i_1)$ for a single ℓ. We index these merged cells by i_1' and denote by $n_\ell(i_1')$ the number of samples in cluster ℓ and cell i_1'. Let the dither distribution satisfy

$$\Pr(\tilde{X} \in \mathscr{C}(i_1') \mid \ell) = \frac{n_\ell(i_1')}{n_\ell} \quad \forall\ i_1'. \qquad (11)$$

Corollary 1. In the 1-dimensional case $d = 1$, Theorem 2 holds if the dither $\tilde{X}$ satisfies Equation (11) for cells i_1' formed by first isolating unique observed values $v_1(i_1)$ and then merging contiguous cells that belong in full to the same cluster, as described above.

Proof. Following the proof of Theorem 2, if $\tilde{x}$ belongs to a cell $\mathscr{C}(i_1')$, then $u = F_{\tilde{X}}(\tilde{x})$ lies in an interval that maps to the set of observed values $v_1(i_1)$ contained in $\mathscr{C}(i_1')$, analogous to Equation (A2). Since Rosenblatt's transformation ensures that u is uniformly distributed, the individual values $v_1(i_1)$ are then selected according to their empirical frequency. □

3.2.2 | Gaussian dither

In this subsection, we assume that the centroid $\bar{x}_\ell$ is the mean of the samples in cluster ℓ, which results when the weighted Euclidean distortion function Equation (4) is chosen. For each cluster ℓ, we also compute the empirical covariance matrix Σ_ℓ from the n_ℓ samples in the cluster. Conditioned on cluster ℓ_i, we then generate $\tilde{x}_i$ as an i.i.d. sample from the Gaussian distribution $\mathcal{N}(\bar{x}_{\ell_i}, \Sigma_{\ell_i} + \alpha I)$. The extra bit of covariance with parameter $\alpha > 0$ accounts for rank-deficient Σ_ℓ, which occurs when the number of distinct values in the cluster (bounded by $n_\ell \geq k$) is less than the dimension d.

The above specification implies that the dither $\tilde{X}$ is distributed as a Gaussian mixture with c mixture components, after marginalizing over the clusters. Closed-form expressions can therefore be given for the first half of Rosenblatt's transformation [8]. Denoted by $\Phi(z; \mu, \sigma^2)$, the Gaussian CDF parameterized by mean μ and variance σ^2, and let $\Lambda_\ell = \Sigma_\ell + \alpha I$. Then for $j = 1$ we have

$$F_{\tilde{X}_1}(\tilde{x}_1) = \sum_{\ell=1}^{c} p_L(\ell) F_{\tilde{X}_1|L}(\tilde{x}_1 \mid \ell) = \sum_{\ell=1}^{c} \frac{n_\ell}{n} \Phi(\tilde{x}_1; \bar{x}_{\ell 1}, \Lambda_{\ell,1,1}), \qquad (12)$$

where $\Lambda_{\ell,j,j}$ is the (j, j) element of Λ_ℓ. For $j > 1$,

$$F_{\tilde{X}_j|\tilde{X}^{j-1}}(\tilde{x}_j \mid \tilde{x}^{j-1}) = \sum_{\ell=1}^{c} p_{L|\tilde{X}^{j-1}}(\ell \mid \tilde{x}^{j-1}) F_{\tilde{X}_j|\tilde{X}^{j-1},L}(\tilde{x}_j \mid \tilde{x}^{j-1}, \ell). \qquad (13)$$

Since $\tilde{X} \mid L$ is multivariate Gaussian, $\tilde{X}_j \mid \tilde{X}^{j-1}, L$ is also Gaussian with parameters given by

$$\mu_{\ell j} = \bar{x}_{\ell j} - \Lambda_{\ell,j,1:j-1}(\Lambda_{\ell,1:j-1,1:j-1})^{-1}(\tilde{x}^{j-1} - \bar{x}_\ell^{j-1})$$
$$\sigma_{\ell j}^2 = \Lambda_{\ell,j,j} - \Lambda_{\ell,j,1:j-1}(\Lambda_{\ell,1:j-1,1:j-1})^{-1}\Lambda_{\ell,1:j-1,j},$$

where $\Lambda_{\ell,j,1:j-1}$ is a row vector consisting of the first $j-1$ columns of the jth row of Λ_ℓ (analogously for $\Lambda_{\ell,1:j-1,j}$), and $\Lambda_{\ell,1:j-1,1:j-1}$ is the submatrix consisting of the first $j-1$ rows and columns of Λ_ℓ. Hence

$$F_{\tilde{X}_j|\tilde{X}^{j-1},L}(\tilde{x}_j \mid \tilde{x}^{j-1}, \ell) = \Phi(\tilde{x}_j; \mu_{\ell j}, \sigma^2_{\ell j}). \tag{14}$$

The conditional probabilities $p_{L|\tilde{X}^{j-1}}(\ell \mid \tilde{x}^{j-1})$ in Equation (13) can be computed recursively over $j=1, \ldots, d$ using Bayes' rule:

$$\begin{aligned} p_{L|\tilde{X}^j}(\ell \mid \tilde{x}^j) &= \frac{p_{L|\tilde{X}^{j-1}}(\ell \mid \tilde{x}^{j-1}) f_{\tilde{X}_j|\tilde{X}^{j-1},L}(\tilde{x}_j \mid \tilde{x}^{j-1}, \ell)}{\sum_{\ell'=1}^{c} p_{L|\tilde{X}^{j-1}}(\ell' \mid \tilde{x}^{j-1}) f_{\tilde{X}_j|\tilde{X}^{j-1},L}(\tilde{x}_j \mid \tilde{x}^{j-1}, \ell')} \\ &= \frac{p_{L|\tilde{X}^{j-1}}(\ell \mid \tilde{x}^{j-1}) \phi(\tilde{x}_j; \mu_{\ell j}, \sigma^2_{\ell j})}{\sum_{\ell'=1}^{c} p_{L|\tilde{X}^{j-1}}(\ell' \mid \tilde{x}^{j-1}) \phi(\tilde{x}_j; \mu_{\ell' j}, \sigma^2_{\ell' j})}, \end{aligned}$$

where the second equality follows from Equation (14) and ϕ denotes a Gaussian probability density function (PDF).

4 | EMPIRICAL RESULTS

In this section, we discuss empirical results obtained using real-world data motivated by the problem of health insurance market risk assessment. Section 4.1 further describes the application. Section 4.2 discusses data sources and simulation of the different populations. Section 4.3 assesses the reidentification risks of the proposed privacy-preservation methods as applied to this data. Section 4.4 presents results on health care expenditure prediction under 3 scenarios: with privacy constraints but without market shift, without privacy constraints but with market shift, and with both privacy constraints and market shift.

4.1 | Market risk assessment

As mentioned in the introduction, market risk assessment involves estimating the health care cost of individuals in a new market who are likely to enroll in an insurer's plan. These estimates can then be appropriately summarized into risk statistics for decision making. Cost information for a new market is typically not available for competitive and other reasons. Instead, insurers have access to cost data for their members in an existing market, plus demographic data for the existing and new markets from public sources, including age, gender, income, veteran status, smoking status, and place of residence. This situation suggests a covariate shift approach as in Section 2 in which a predictive model relating demographic variables, X, to health care cost, Y, is learned from the existing market population and is then applied to the new market's demographic data, taking into account the different demographic distributions in the existing and new markets, p_X and q_X. Since the cost Y is regarded as being continuous, the learning problem (Equation (1)) is one of the regression. Any regression technique can be used for this purpose, in particular those that account for the skewness and heteroscedasticity of health care costs, for example ordinary least-squares with log-transformed data, 2-part models, generalized linear models, and multiplicative regression [6,12,37]. These correspond to different choices for the function class $\mathcal{F}$ and loss function $\mathcal{L}$ in Equation (1).

However, market risk assessment differs from the standard covariate shift setting in Section 2 in having a distinction between member populations and larger market populations, since only a subset of the latter enroll in a given health insurance plan. 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 market from the new market ($M=1$ means new market). Since training data with costs comes from the insurance company's data on current plan members, the training distribution, previously denoted as p_X in Section 2, is now denoted as $p_{X|E,M}(x|e=1, m=0)$, referring to enrollees in the current market. Likewise, the test distribution q_X is $p_{X|E,M}(x|e=1, m=1)$ for enrollees in the new market.

In some cases it may be possible to estimate $p_{X|E,M}(x|1, 1)$ 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. The more general and usual case is that $p_{X|E,M}(x|1, 1)$ cannot be estimated directly. In this scenario, besides the current member distribution $p_{X|E,M}[x|1, 0]$, we may still assume that demographic distributions for the current and new markets are available, corresponding to $p_{X|M}(x|0)$ and $p_{X|M}(x|1)$, respectively. These distributions are related by Bayes' rule,

$$p_{X|E,M}(x \mid 1, m) = \frac{p_{E|X,M}(1 \mid x, m) p_{X|M}(x \mid m)}{p_{E|M}(1 \mid m)}, \qquad m = 0, 1. \tag{15}$$

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

$$\frac{p_{X|E,M}(x \mid 1, 1)}{p_{X|E,M}(x \mid 1, 0)} \propto \frac{p_{E|X,M}(1 \mid x, 1)}{p_{E|X,M}(1 \mid x, 0)} \frac{p_{X|M}(x \mid 1)}{p_{X|M}(x \mid 0)} \tag{16}$$

as functions of x.

In this work, we make the assumption that $p_{E|X,M}(1|x, m)$, 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 $p_{E|X,M}(1|x, m) = p_{E|X}(1|x)$. This is a reasonable starting assumption positing that enrollment depends on demographic variables such as age, sex, etc., but does not depend on which market the individual belongs to once those demographic variables are specified. This assumption may be later modified by an insurance market expert to account for additional market factors such as the level of competition.

With the conditional independence assumption, Equation (16) 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 $p_{X|E,\,M}(x|1,\ 0)$ while the test samples are distributed according to $p_{X|E,\,M}(x|1,\ 1)$, the importance weighting is therefore $w(x)\ =\ p_{X|M}(x|1)/p_{X|M}(x|0)$, taking the place of $w(x) = q_X(x)/p_X(x)$ in Section 2. The importance weights may be estimated in the same manner as before, for example, nonparametrically using empirical distributions, or using logistic regression. We evaluate both of these covariate shift methods in Sections 4.4.2 and 4.4.3.

A second notable feature of market risk assessment, and insurance applications in general, is the emphasis on aggregate predictions of the average or total cost for groups of individuals rather than individual-level predictions. For example, one may be interested in the average cost for a new enrollee population as a whole or for segments of the population. Given a regression model constructed as described above, aggregate predictions can be obtained simply by averaging the predictions for each individual in the group. In terms of prediction error, this averaging has the effect of greatly attenuating the error variance: For a group of m i.i.d. individuals, the error variance decreases by a factor of m. As a consequence, for large m the bias of the predictor, $b(\widehat{Y}) = \mathbb{E}[\widehat{Y}] - \mathbb{E}[Y]$, becomes the dominant measure of error as compared to variance or R^2, a common representation of individual-level MSE. We refer the reader to Equation (8) for a straightforward calculation decomposing aggregate prediction error into bias and variance/R^2 components. Thus in reporting prediction results in Section 4.4, we focus more on bias performance rather than variance or R^2.

4.2 | Description of data

We use publicly available Medical Expenditure Panel Survey (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 produced by the United States Department of Health and Human Services' Agency for Healthcare Research and Quality, 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 United States 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, $p_{X|M}(x|0)$, can be taken to be that of the United States, while the new market distributions $p_{X|M}(x|1)$ correspond to California rating areas. To simulate these two markets, the MEPS data set is randomly and evenly split into training and test sets. All results reported in Sections 4.4.2-4.4.3 are averaged over 200 such splits. The existing market distribution $p_{X|M}(x|0)$ is estimated empirically directly from the training set. The distribution $p_{X|M}(x|1)$ 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) [2].

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 $p_{X|E,\,M}(x|1, 0)$, samples in the existing market data set are reweighed based on the age distribution in the initial enrollment period of the Health Insurance Marketplaces created by the Affordable Care Act [3,36], again relative to the national baseline. The resulting distribution differs notably from that of the larger market, $p_{X|M}(x|0)$, in having few children (< 18) and seniors (> 65). The age-dependent enrollment probabilities $p_{E|X}(1|x)$ induced by this procedure are then applied to samples in the new market to simulate plan enrollment in the new market, $p_{X|E,\,M}(x|1,\ 1)$.

We also consider a second, simpler scenario where the demographic distribution of the existing and the new markets, $p_{X|M}(x|0)$ and $p_{X|M}(x|1)$ are both taken to be that of the United States. The MEPS data set is randomly and evenly split into training and test sets, but the test set is not reweighted in this case. We present results for this scenario in Section 4.4.1 for 200 such splits. The plan distributions are simulated using the same procedure described above and hence are also equal to each other.

The specific MEPS data set we consider is for the year 2005, containing just over 15 000 weighted records, and the demographic variables are gender, age (binned into 8 groups similar to those in Ref. [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.

4.3 | Reidentification risk

In this section, we empirically evaluate the reidentification risks of the privacy-preservation methods proposed in Section 3. Here reidentification risk refers to the probability of correctly retrieving an individual's record from an anonymized data set using their true quasi-identifiers x. It is seen that for a given anonymity parameter k, the distribution-recovering transformations discussed in Section 3.2 have reidentification risks at least as good as or better than standard k-anonymization.

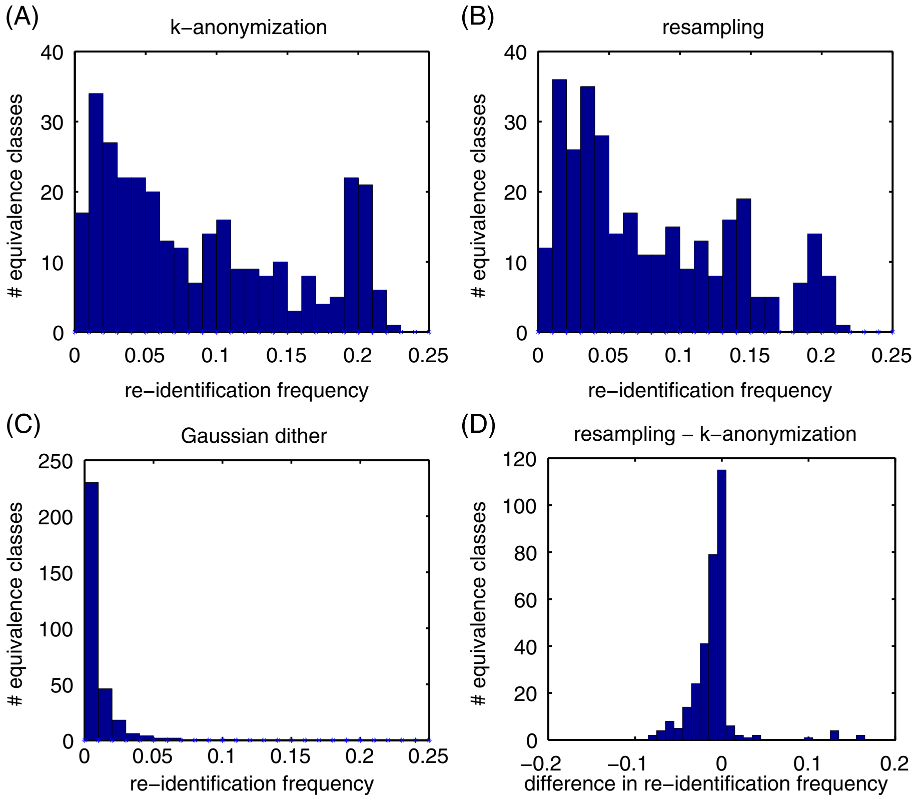

FIGURE 3 Histograms of reidentification frequencies over 1000 trials for anonymity parameter $k = 5$. The proposed distribution-preserving method using (B) resampling with replacement has similar reidentification risk as (A) standard k-anonymization, while the distribution-preserving method using (C) Gaussian dither has substantially lower risk. In (D), a histogram of differences in reidentification frequencies between the resampling method and k-anonymization indicates that the former is better for a large majority of equivalence classes

Enrollment data for the existing market (corresponding to $p_{X|E,M}(x|1, 0)$) is simulated from MEPS as described in Section 4.2. Recall that this is the data set that requires privacy protection because it contains health care expenditures. Each record in the data set is assigned a record number. The data set then undergoes one of three procedures: (1) standard k-anonymization through clustering (Section 3.1), (2) clustering plus distribution recovery via resampling with replacement (Section 3.2.1), and (3) clustering plus distribution recovery via Gaussian dither (Section 3.2.2). The result is an anonymized data set with modified quasi-identifiers $\hat{x}$. Before clustering, all variables x and y are standardized. The means and variances are restored after the privacy transformations. We set $w = 1$ in the distortion function (4) for clustering and add diagonal loading $\alpha = 1/3$ to the Gaussian dither.

For reidentification, each record in the original data set is matched to the anonymized data set based on quasi-identifiers, x and $\hat{x}$, respectively. A minimum Euclidean distance criterion is used so that a match is always returned, even if inexact ($\hat{x} \neq x$). In the frequent case of multiple matches (multiple $\hat{x}$ at the minimum distance), one of the anonymized records is selected uniformly at random. A reidentification is declared if the record numbers of the original and matched records are the same.

We first discuss the reidentification performance to be expected under the above scheme. Successful reidentification requires two events to occur: (1) the original quasi-identifiers x are mapped to a minimum-distance point $\hat{x}$ (often but not necessarily identical to x), and (2) the correct record is selected from multiple matches. Conventional k-anonymization focuses on limiting the probability of the second event to no more than $1/k$ by ensuring that every combination of quasi-identifier values (termed an equivalence class) in the anonymized data set is shared by at least k records. The probability can be less than $1/k$ if some equivalence classes have more than k records. In contrast, for the proposed distribution-preserving methods, the distribution over equivalence classes follows that of the original data set and thus cannot be expected to satisfy the k-record requirement. Instead it is the probability of the intersection of the two events that is controlled at the same $1/k$ level. While an analysis of the general case is not straightforward, for the case of resampling with replacement within a cluster of size k with no values in common with other clusters, it can be verified that the probability of the intersection is exactly $1/k$, as expected from symmetry. If there are common values, then the probability is generally reduced because the second event of correct selection becomes less likely if a common value is matched. In the case of Gaussian dither, the first event becomes less likely because there is some probability of mapping x to a more distant point outside the cluster.

With this intuition in mind, we now discuss empirical results obtained by averaging multiple trials of the above anonymization and reidentification procedure. The

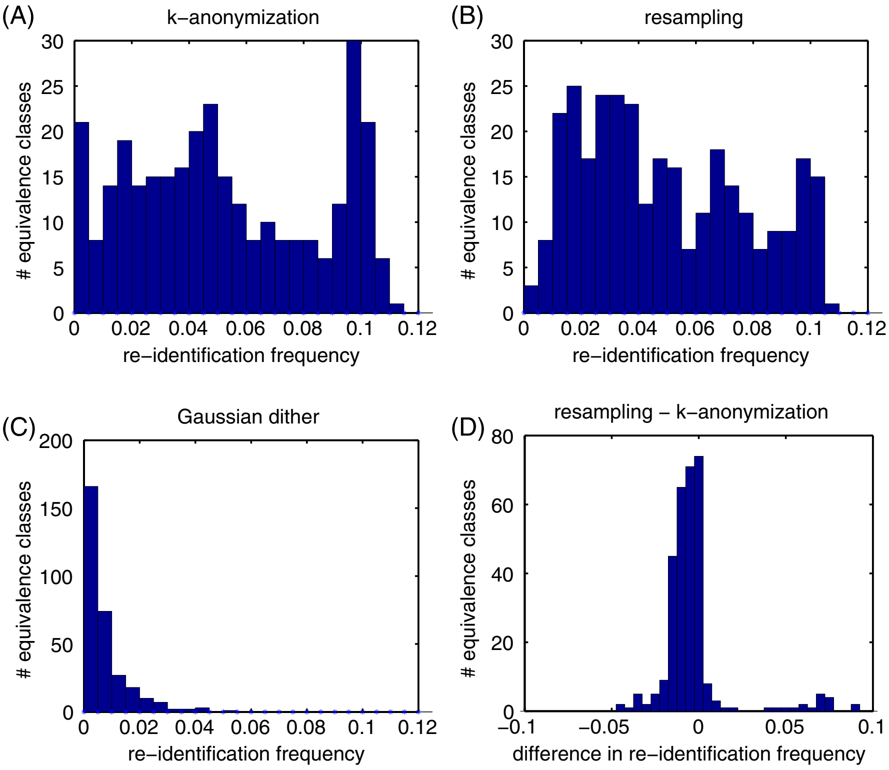

FIGURE 4 Histograms of reidentification frequencies over 2000 trials for anonymity parameter $k = 10$

randomness in each trial is due to selection from multiple matches in reidentification as well as dither in the distribution-preserving methods. Since the reidentification algorithm depends on a record only through the equivalence class of x, we report reidentification frequencies by equivalence class, of which there are 310 in the original data set.

Figure 3 shows histograms of reidentification frequencies over equivalence classes for the 3 privacy protection methods, k = 5, and 1000 trials. For standard k-anonymization in Figure 3A, the reidentification frequency ranges from 0 to approximately $1/k$ = 1/5. As explained above, the distribution reflects the fact that many equivalence classes in the original data set already exceed 5 members, in some cases by a large factor. Similarly large equivalence classes are maintained after clustering. The few reidentification frequencies greater than 1/5 can be attributed to sampling error. In the worst case, the reidentification frequency is proportional to a binomial random variable with parameters 1000 and 1/5, and hence the observed frequency may exceed 1/5.

For the distribution-preserving method using resampling, Figure 3B shows that the reidentification risk is similar to that of k-anonymization. The histogram depends again on the range of equivalence class sizes and also on the prevalence of common values between clusters, as mentioned earlier. On the other hand, the Gaussian dither method in Figure 3C has much lower reidentification risk. This can be attributed to the choice of diagonal loading $\alpha = 1/3$, which causes the Gaussian dither to frequently map original quasi-identifiers x to points $\hat{x}$ that are beyond minimum distance. Reducing α would increase the reidentification risk. Figure 3D provides a more detailed comparison between the resampling method and k-anonymization by plotting the histogram of differences in reidentification frequencies. For a large majority of equivalence classes and $k = 5$, resampling offers slightly better privacy protection than k-anonymization. This is partly offset by worse protection for a few equivalence classes. Figure 4 shows that all of the above patterns hold for k = 10 and 2000 trials.

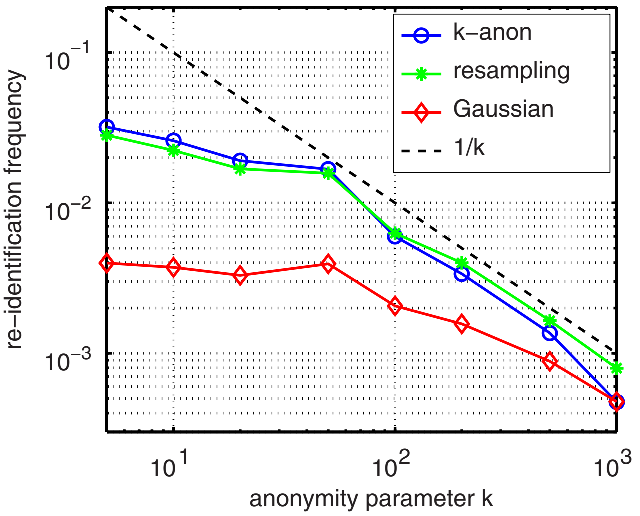

FIGURE 5 Reidentification frequency, averaged over all records and 200 trials, for different privacy protection methods as a function of anonymity parameter k

For larger k, the reidentification probabilities decrease accordingly, and direct empirical validation on a per-equivalence-class basis would require increasing numbers

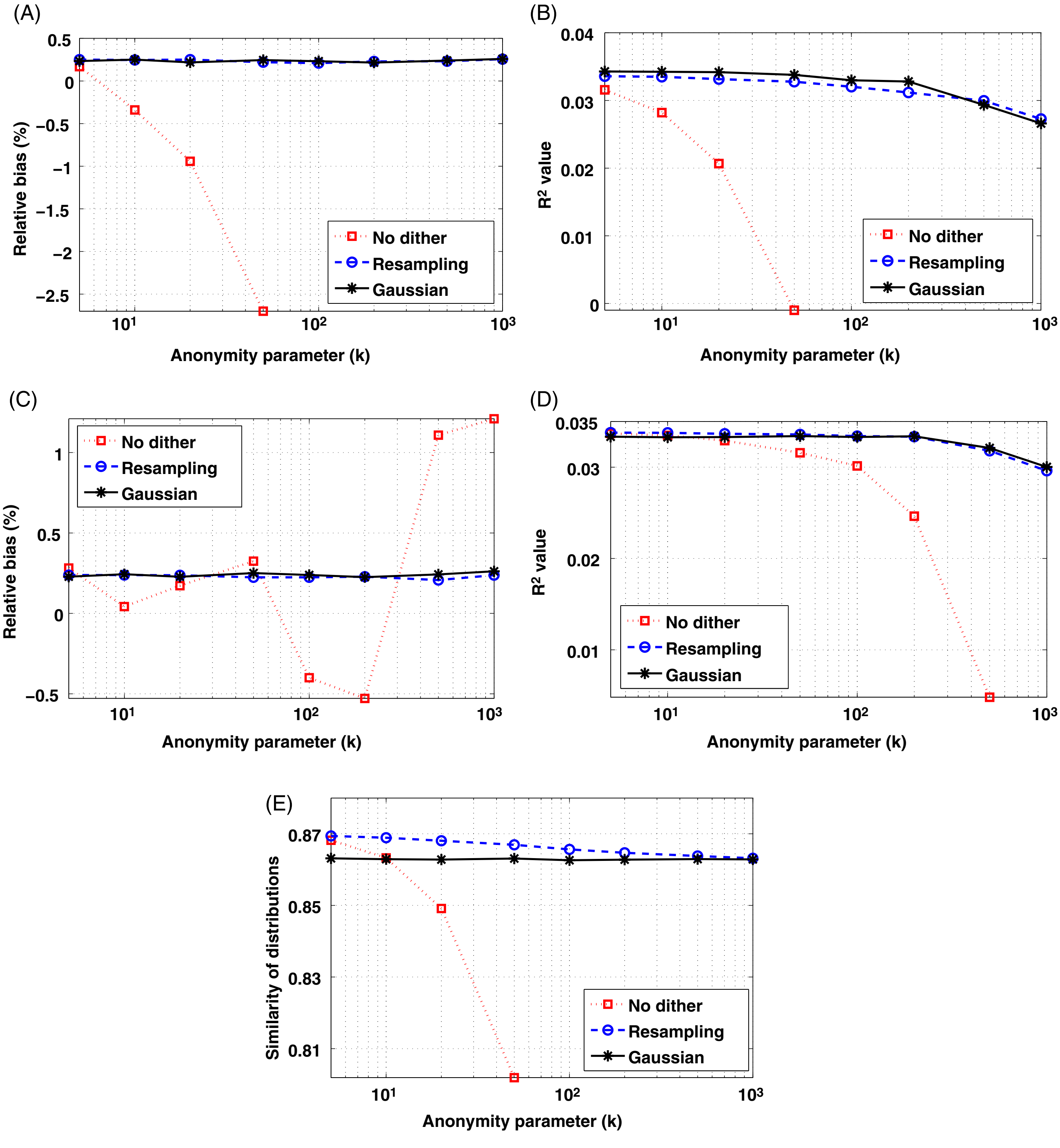

FIGURE 6 A, Prediction bias for dummy-coded variables. B, R^2 coefficient for dummy-coded variables. C, Prediction bias for numeric variables. D, R^2 coefficient for numeric variables. E, Histogram similarity of distributions resulting from k-anonymization with and without the proposed distribution-preserving procedures. For k-anonymity without distribution preservation, relative bias in (A), R^2 in (B) and (D), and distribution similarity in (E) all drop drastically with k and reach minimum values of −53.2, − 1.193, − 0.0139, and 0.1709, respectively (not shown) for $k = 1000$. In (C), the lack of distribution preservation makes the bias oscillate but it remains relatively small for all values of k. In contrast, both distribution-preserving procedures result in almost constant bias and only small drops in R^2 and distribution similarity

of trials to combat the binomial sampling error mentioned above. As an alternative, in Figure 5 we summarize performance by plotting the average reidentification frequency over all records and 200 trials as a function of k. In all cases, these average reidentification frequencies fall below the nominal value of $1/k$ due to the respective reasons given earlier in this section. The margin is most substantial for $k < 50$ because of large equivalence classes. For smaller k, the resampling method happens to yield slightly better privacy than k-anonymization while the Gaussian dither method is much stronger, in agreement with Figures 3 and 4. As k increases, k-anonymization and the resampling method exchange places while the advantage of the Gaussian method diminishes.

TABLE 1 Dummy-coded variable performance results for rating areas of California as new markets for no shift, logistic shift, and nonparametric shift

	R^2 value			Relative bias (%)		
New market	No shift	Logistic	Nonparametric	No shift	Logistic	Nonparametric
RA 1	0.0300	0.0271	0.0280	4.38	4.64	1.86
RA 2	0.0243	0.0228	0.0220	3.75	2.14	−0.49
RA 3	0.0289	0.0272	0.0270	2.34	1.87	−0.24
RA 4	0.0291	0.0278	0.0264	−2.54	−2.39	1.40
RA 5	0.0221	0.0212	0.0214	3.63	1.76	0.53
RA 6	0.0235	0.0232	0.0227	2.17	1.42	0.69
RA 7	0.0233	0.0227	0.0219	2.31	1.40	−0.39
RA 8	0.0213	0.0213	0.0196	1.12	−0.69	−1.29
RA 9	0.0233	0.0219	0.0230	5.99	5.67	0.32
RA 10	0.0330	0.0316	0.0323	3.48	2.87	0.57
RA 11	0.0314	0.0290	0.0306	4.90	4.87	2.17
RA 12	0.0246	0.0233	0.0243	4.29	3.49	0.18
RA 13	0.0328	0.0294	0.0253	−1.03	−0.53	0.10
RA 14	0.0345	0.0331	0.0330	2.47	1.79	0.36
RA 15-16	0.0295	0.0281	0.0286	2.83	2.62	0.81
RA 17	0.0318	0.0300	0.0302	0.72	0.24	−0.77
RA 18	0.0250	0.0246	0.0252	3.75	2.85	0.84
RA 19	0.0268	0.0258	0.0268	3.36	3.07	1.23

TABLE 2 Numeric variable performance results for rating areas of California as new markets for no shift, logistic shift, and nonparametric shift

	R^2 value			Relative bias (%)		
New market	No shift	Logistic	Nonparametric	No shift	Logistic	Nonparametric
RA 1	0.0312	0.0286	0.0291	0.47	4.49	1.81
RA 2	0.0260	0.0247	0.0242	1.15	1.76	−0.48
RA 3	0.0288	0.0275	0.0273	0.06	1.56	−0.30
RA 4	0.0296	0.0289	0.0279	−2.39	−1.94	1.27
RA 5	0.0247	0.0238	0.0238	0.97	1.47	0.48
RA 6	0.0252	0.0249	0.0248	0.22	1.43	0.63
RA 7	0.0260	0.0257	0.0256	0.90	1.65	−0.45
RA 8	0.0245	0.0239	0.0232	−1.20	−0.83	−1.34
RA 9	0.0236	0.0225	0.0235	4.41	5.94	0.30
RA 10	0.0316	0.0304	0.0311	3.01	3.57	0.50
RA 11	0.0308	0.0290	0.0305	3.16	5.67	2.04
RA 12	0.0265	0.0253	0.0259	1.15	3.41	0.13
RA 13	0.0326	0.0307	0.0291	−0.85	0.13	−0.09
RA 14	0.0337	0.0328	0.0328	1.87	2.52	0.20
RA 15-16	0.0292	0.0287	0.0290	2.96	3.05	0.73
RA 17	0.0312	0.0301	0.0300	0.86	0.86	−0.85
RA 18	0.0258	0.0252	0.0258	2.63	3.00	0.79
RA 19	0.0283	0.0272	0.0279	1.45	3.18	1.17

4.4 | Privacy preservation and market shift

We now discuss results in predicting health care expenditure as a function of the demographic variables age, gender, education level, and income level, as detailed in Section 4.2. Age, education, and income level are ordinal variables while gender is binary. The regression model used in all cases is a sum of univariate functions of each demographic variable. We consider 2 cases as follows.

1. The univariate functions are not constrained to be linear and can vary arbitrarily with input value. We refer to this case as “dummy-coded” since it can be achieved by dummy-coding the discrete variables, that is, introducing

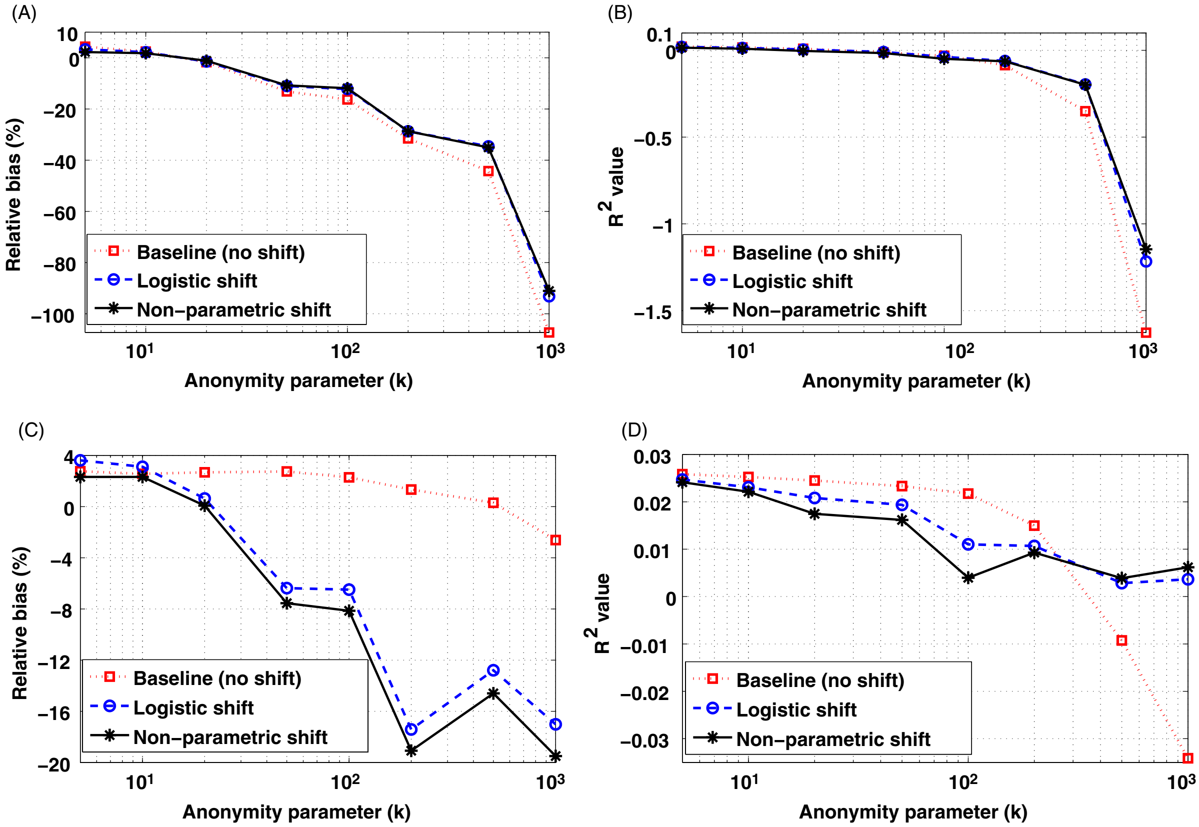

FIGURE 7 Performance results with no distribution preservation: A, Dummy-coded prediction bias. B, Dummy-coded R^2 coefficient. C, Numeric prediction bias. D, numeric R^2 coefficient. For dummy-coded data, the prediction error increases unacceptably as k increases. For numeric data, the bias drops only moderately for the baseline case whereas the R^2 coefficient becomes unacceptable for large values of k. With the covariate shift methods, the bias drops rapidly, whereas the R^2 coefficient decreases slowly, as k increases

a binary variable for each level while leaving 1 level out to avoid linear dependence on the intercept term.

2. The univariate functions are constrained to be linear in each variable. We refer to this second case as "numeric" since it is achieved by interpreting the levels (e.g. 0-7 for age, 0-5 for education) as real numeric values.

The regression model in the second case has many fewer degrees of freedom than the more flexible model of the first case. For this specific MEPS data set, it can be observed that the first model does not offer improvements in prediction accuracy compared to the simpler second model. We include results for the first model to illustrate the behavior of a commonly used, more complicated model when subjected to privacy preservation and covariate shift.

4.4.1 | Prediction with privacy preservation

We first present results for the simpler scenario where there is no shift in the demographic distributions between the existing and the new markets, as well as between the corresponding plan enrollment distributions. The plan enrollment data from the existing market, that is, the training data, is subjected to the same three types of privacy-preserving transformations as in Section 4.3. Figure 6 shows the relative bias and coefficient of determination R^2 in predicting health care cost in the new market as the anonymity parameter k increases. For the nonlinear dummy-coded model in Figure 6A and Figure 6B, it is clear that performance drops dramatically when distribution-preserving approaches are not used. Conversely, performance remains either constant or only drops slightly using the proposed distribution-preserving approaches. For the linear numeric model in Figure 6C and Figure 6D, the difference is less stark: Without distribution preservation, relative bias oscillates at a relatively low level while R^2 decreases to a negative but less extreme value. These results suggest that the simpler linear model is inherently more robust to privacy transformations but can nevertheless be improved by the proposed distribution-preservation methods.

The poor performance of k-anonymization without distribution preservation can be understood from two perspectives. First, since in conventional k-anonymization, samples of X are replaced with the corresponding cluster centers $\bar{x}_\ell$, the number of unique samples for training the regression model decreases to $\left\lfloor \frac{n}{k} \right\rfloor$, a substantial reduction when k is large. It is well-known that the prediction error of a linear regression model tends to decrease at a $O(1/\sqrt{n})$ rate, where n is the training sample size [32]. Second, in Figure 6E we plot the similarity between the training data distribution after privacy transformation and the test data distribution. We use the histogram intersection similarity [34] for concreteness, for which a value close to 1 implies that the predictor is trained

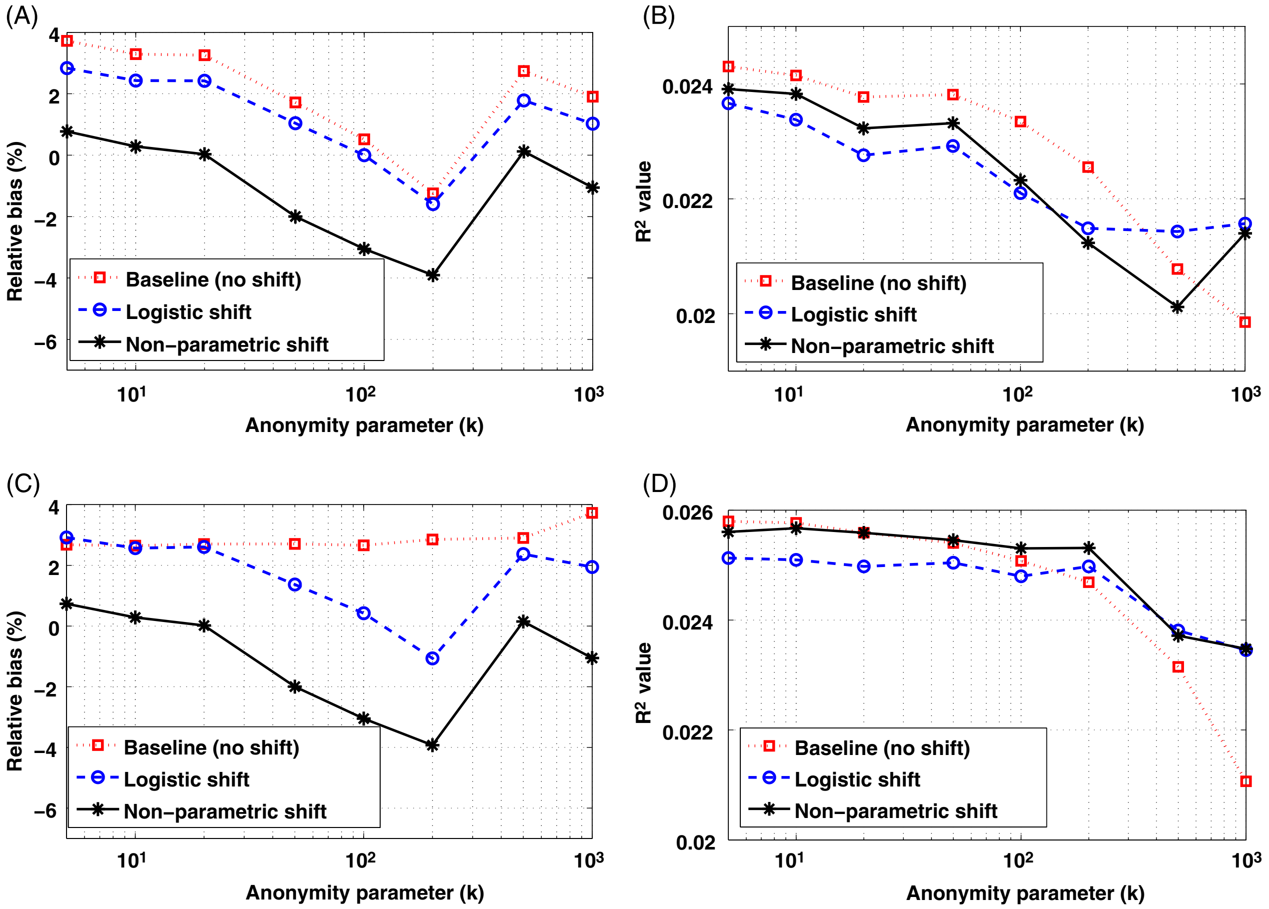

FIGURE 8 Performance results for intra-cluster dither. A, Dummy-coded prediction bias. B, Dummy-coded R^2 coefficient. C, Numeric prediction bias. D, Numeric R^2 coefficient. For dummy-coded data, as k increases, relative bias stays low whereas the R^2 coefficient has a generally decreasing trend. For numeric data, as k increases, the baseline results in an almost-constant bias and a steady decrease in the R^2 coefficient. For the 2 shift methods, the relative bias with numeric data is similar to the case with dummy-coded data

on a distribution much like the one encountered in testing. Figure 6E indicates that this similarity falls dramatically with k under conventional k-anonymization. Overall, it is clear that distribution preservation is beneficial to achieving useful k-anonymity even when simple workloads such as linear regression are involved.

4.4.2 | Prediction with market shift

Next we discuss cost prediction in the absence of privacy-preserving data transformations but with a shift in demographic distribution between the existing and the new markets. Two covariate shift methods are compared, corresponding respectively to nonparametric and logistic regression methods of estimating the importance weights $w(x)$. A baseline method that does not account for covariate shift is also compared.

Tables 1 and 2 summarize performance in predicting the cost in rating areas of California as new markets using nonlinear dummy-coded and linear numeric models respectively. As discussed in Section 4.1, for aggregate prediction the bias is often the more important performance metric. Table 1 shows that the baseline approach using the dummy-coded model has a noticeable bias, and the covariate shift approaches reduce the bias for most new markets 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 nonparametric shift method. With the linear numeric model in Table 2, the relative bias is lower for the baseline, but still, the nonparametric shift method tends to offer a reduction over the baseline approach in most rating areas. In fact, the biases of the covariate shift methods under the dummy-coded model and the numeric model are similar in most markets. Performing covariate shift is more useful with the nonlinear dummy-coded model, which is very common in several applications, but also provides some advantage for the numeric model.

4.4.3 | Prediction with privacy preservation and market shift

Lastly we present results for the scenario in which both the insurer's existing plan data is subjected to the three privacy transformations used in Sections 4.3 and 4.4.1, and there is a demographic shift between the existing and new markets. We focus in this subsection on California rating area 18 as the new market.

Figure 7 shows results for k-anonymization without distribution preservation. As k increases, the original samples from

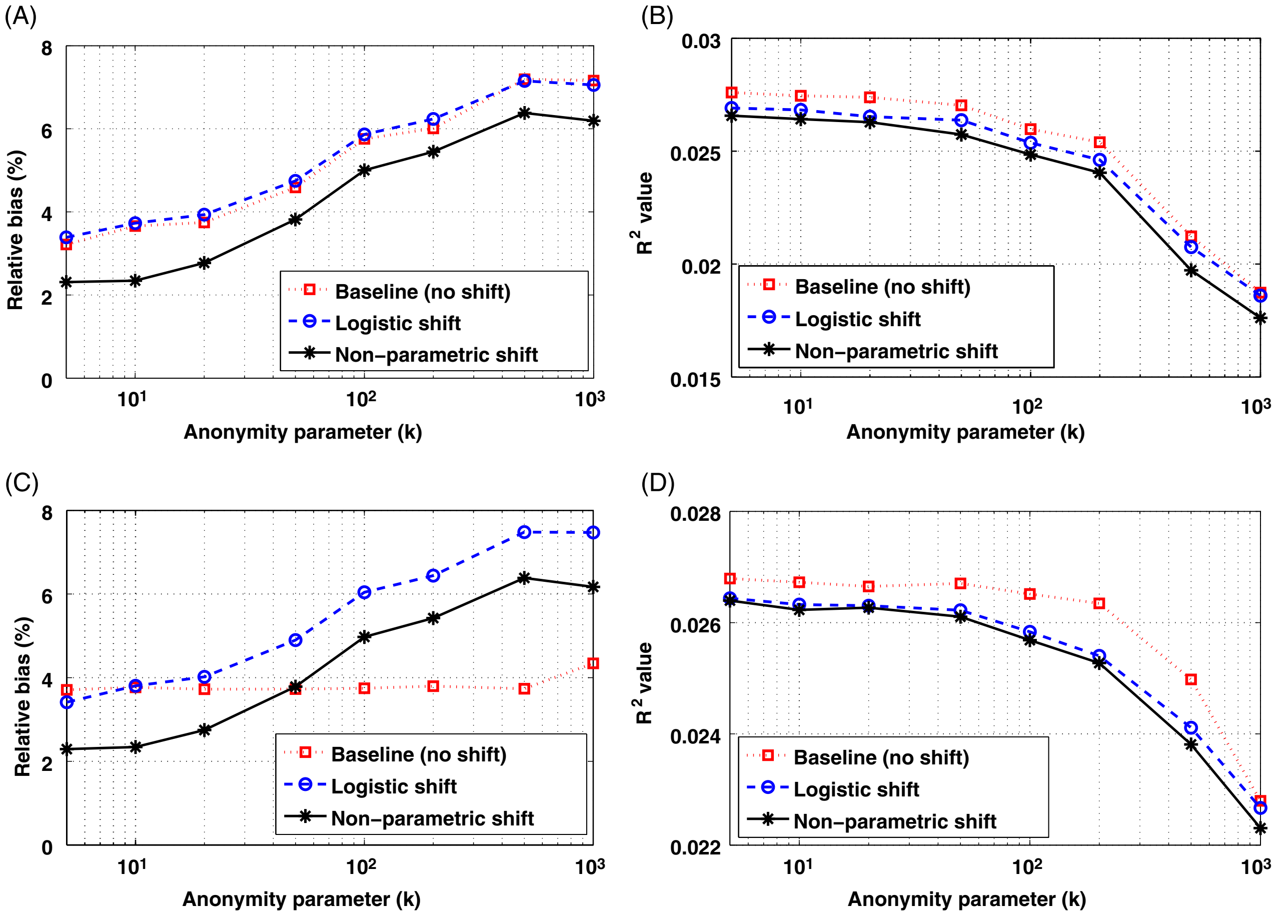

FIGURE 9 Performance results for Gaussian dither. A, Dummy-coded prediction bias. B, Dummy-coded R^2 coefficient. C, Numeric prediction bias. D, Numeric R^2 coefficient. For dummy-coded data, as k increases, distribution preservation moderates the increase in bias, while the covariate shift methods reduce bias. For numeric data, as k increases, the baseline results in an almost-constant bias and a steady decrease in the R^2 coefficient. For the 2 shift methods, the relative bias with numeric data is similar to the case with dummy-coded data

the plan data are represented more and more coarsely by their cluster centers. As a consequence, covariate shift methods fail and prediction accuracy suffers markedly. In fact, both bias and R^2 are unacceptably bad for the dummy-coded data model. For the numeric data model, the bias does not actually deteriorate by much under the baseline approach, again pointing toward the robustness of the simple linear model. However, R^2 still decreases to negative values. The covariate shift methods continue to yield unacceptable bias but manage to keep R^2 positive.

Figure 8 shows the prediction performance of the proposed privacy-preserving procedure using resampling with replacement combined with different covariate shift methods. In great contrast to Figure 7, the relative bias stays low for all values of k while R^2 decreases only slightly with increasing k. The main difference between the dummy-coded and numeric models is that the bias remains almost constant in the latter case when the baseline method is used.

Figure 9 depicts performance under privacy preservation with Gaussian dither. In this case, the bias remains controlled but increases with k to a level higher than for the resampling method in Figure 8. A possible explanation for the difference between the 2 distribution-preserving methods can be found in Figures 3–5 and Section 4.3. There it is seen that the Gaussian dither method has much lower reidentification risk, likely as a result of mapping original quasi-identifiers x to distant points $\hat{x}$. This mapping however may further distort the relationship between X and Y relative to the original, causing the prediction bias to increase. In the dummy-coded case, the nonparametric covariate shift method succeeds at reducing bias for all k. In the numeric case, covariate shift methods reduce bias only for $k < 50$. 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.

Overall, Figures 7–9 demonstrate the advantage of preserving quasi-identifier distributions when k-anonymization is required in a covariate shift setting. Some insight into the above results can be seen in Figure 10, which depicts the same histogram intersection similarity used in Figure 6E, this time between the new enrollment distribution estimated from privacy-transformed training data using the covariate shift methods, and the actual new enrollment. Under conventional k-anonymization in Figure 10A, the similarity decreases rapidly with k, but using distribution-preserving privacy transformations in Figure 10B and Figure 10C, the

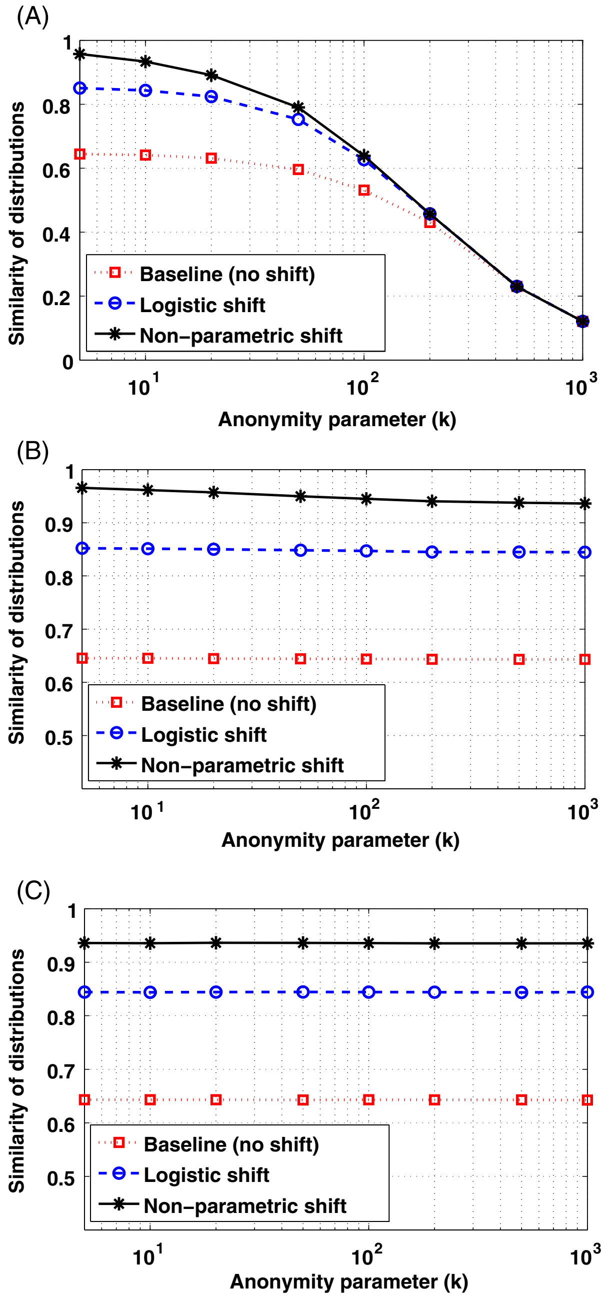

FIGURE 10 Area under histogram intersection between the estimated and actual new enrollment distributions for different covariate shift methods and (A) no distribution preservation, (B) intra-cluster dither, or (C) Gaussian dither. Both proposed distribution-preserving transformations maintain a mostly constant similarity as the anonymity k increases, whereas the similarity deteriorates rapidly without distribution preservation

similarity can be kept mostly constant as the anonymity k increases and can be further enhanced by the covariate shift methods.

5 | CONCLUSION

In this paper, our main contribution has been to develop a new privacy-preservation operation that has the key property of preserving the data distribution of the quasi-identifiers. The specific privacy criterion we address is k-anonymity, which is a common mathematical interpretation of legal privacy standards. We have shown how such distribution preservation is a clear need in supervised machine learning settings such as covariate shift and transfer learning that also require anonymized microdata. Other data analysis workloads that require both anonymization and lack of distortion in the data distribution will also benefit from the proposed methodology. Workloads requiring distribution preservation have not yet been addressed in the data privacy literature.

Our proposed technique combines k-member clustering with components from distribution-preserving quantization, namely dithering and Rosenblatt's transformation. Distribution-preserving quantization was originally developed in the audio signal processing literature and has never been applied to privacy applications before. We have analyzed two approaches for dithering, intra-cluster and Gaussian, and shown how both achieve distribution preservation when followed by Rosenblatt's transformation. Existing distribution-preserving quantization takes the number of clusters as input rather than constraining the cluster size; thus, another contribution of our work is the extension of distribution-preserving quantization to a constrained setting.

Moreover, we have contributed to a solution to the real-world health care market risk assessment problem, a common problem encountered by health insurance companies that was especially pertinent after passage of the Affordable Care Act. Insurance companies had not developed machine learning approaches for this problem, only relying on crude estimates mainly driven by intuition and coarse aggregate-level data. We show successful empirical results on MEPS data with realistic simulations for new markets and enrollment probabilities. (Actual health cost data from insurance companies that we have worked with in the recent past is confidential.)

In particular, we see that the overall method with intra-cluster dithering keeps the reidentification risk approximately the same as k-member clustering, which is sufficient for the requirement of k-anonymity. The Gaussian dither has even lower reidentification risk. Additionally, the nonparametric version of the covariate shift is successful in significantly reducing the relative bias of the regression that would occur if the covariate shift were not done. Examining the results of the full solution, we see that without our new privacy-preservation methods, cost prediction can fail for k-anonymity greater than 10 or 20, but prediction results remain satisfactory when using our proposed approach.

One health care application-specific 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, that is, 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 $\mathbb{E}[Y \mid X, M] = a(M)\mathbb{E}[Y \mid X]$).

A general 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. We would also like to explore stronger notions of privacy such as l-diversity [24] and t-closeness [21] as well as nondiscrimination and fairness (an issue that is mathematically similar to privacy [30]), all of which have not been examined for distribution-preserving workloads. Exploration of different solution methods for the k-member clustering optimization problem is of interest as well.

ACKNOWLEDGMENT

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

ORCID

Kush R. Varshney http://orcid.org/0000-0002-7376-5536

REFERENCES

1. M. Alamgir, G. Lugosi, and U. von Luxburg, *Density-preserving quantization with application to graph downsampling*, in *Proceedings of the Conference on Learning Theory (Barcelona, Spain, 2014)*, 2014, 543–559, available at http://proceedings.mlr.press/v35/.
2. American Community Survey, United States Census Bureau, 2005, available at http://www.census.gov/acs/www/data_documentation/pums_data.
3. ASPE "Health insurance marketplace: Summary enrollment report for the initial annual open enrollment," United States Department of Health and Human Services, ASPE Issue Brief, May 2014.
4. M.-F. Balcan, A. Blum, S. Fine, and Y. Mansour, "*Distributed learning, communication complexity and privacy*," in *Proceedings of the Conference on Learning Theory (Edinburgh, UK, 2012)*, 2012. 26, available at http://proceedings.mlr.press/v23/.
5. A. Banerjee and J. Ghosh, *Scalable clustering algorithms with balancing constraints*, Data Min. Knowl. Disc. 13 (2006), 365–395.
6. A. Basu and W. G. Manning, *Issues for the next generation of health care cost analyses*, Med. Care 47 (2009), S109–S114.
7. S. Basu, I. Davidson, and K. Wagstaff, *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, CRC Press, Boca Raton, FL, 2009.
8. S. Bickel, C. Sawade, and T. Scheffer, *Transfer learning by distribution matching for targeted advertising*, Adv. Neur. Inf. Process. Syst. 21 (2009), 145–152.
9. J.-W. Byun et al., *Efficient* k-*anonymization using clustering techniques*, in *Proceedings of the International Conference Database Systems for Advanced Applications (Bangkok, Thailand, 2007)*, 2007, 188–200, available at https://link.springer.com/book/10.1007/978-3-540-71703-4.
10. J. P. Daries et al., *Privacy, anonymity, and big data in the social sciences*, Commun. ACM 57 (2014), 56–63.
11. A. Demiriz, K. P. Bennett, and P. S. Bradley, *Using assignment constraints to avoid empty clusters in k-means clustering*, in *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, S. Basu, I. Davidson, and K. Wagstaff, Eds., CRC Press, Boca Raton, FL, 2009, 201–220.
12. P. Diehr et al., *Methods for analyzing health care utilization and costs*, Annu. Rev. Public Health 20 (1999), 125–144.
13. K. El Emam and F. K. Dankar, *Protecting privacy using k-anonymity*, J. Am. Med. Inform. Assoc. 157 (2008), 627–637.
14. N. Ganganath, C.-T. Cheng, and C. K. Tse, *Data clustering with cluster size constraints using a modified* k-*means algorithm*, in *Proceedings on the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (Shanghai, China, 2014)*, IEEE Computer Society, 2014, 158–161.
15. R. Ge et al., *Constraint-driven clustering*, in *Proceedings of ACM SIGKDD International Conference on. Knowledge Discovery and Data Mining (San Jose, CA, 2007)*, ACM, 2007, 320–329.
16. S. Geetha, G. Poonthalir, and P. T. Vanathi, *Improved k-means algorithm for capacitated clustering problem*, INFOCOMP J. Comp. Sci. 8 (2009), 52–59.
17. R. M. Gray and D. L. Neuhoff, *Quantization*, IEEE Trans. Inf. Theory 44 (1998), 2325–2383.
18. V. S. Iyengar, Transforming data to satisfy privacy constraints, in *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Edmonton, Canada, 2002)*, ACM, 2002, 279–288.
19. H. Kargupta et al., *On the privacy preserving properties of random data perturpation techniques*, in *Proceedings of the IEEE International Conference on Data Mining (Melbourne, FL, 2003)*, IEEE, 2003, 99–106.
20. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, *Mondrian multidimensional k-anonymity*, in *Proceedings of the IEEE International Conference on Data Engineering (Atlanta, GA, 2006)*, IEEE, 2006.
21. N. Li, T. Li, and S. Venkatasubramanian, *t-closeness: Privacy beyond k-anonymity and l-diversity*, in *Proceedings of the IEEE International Conference on Data Engineering (Istanbul, Turkey, 2007)*, IEEE, 2007, 106–115.
22. M. Li, J. Klejsa, and W. B. Kleijn, *Distribution preserving quantization with dithering and transformation*, IEEE Signal Process. Lett. 17 (2010), 1014–1017.
23. S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, *Quantization and dither: A theoretical survey*, J. Audio Eng. Soc. 40 (1992), 355–375.
24. A. Machanavajjhala et al., *l-diversity: Privacy beyond k-anonymity*, ACM Trans. Knowl. Discov. Data 1 (2007), 3.
25. 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 (Jan. 2011), 3–10.
26. D. G. Messerschmitt, *Quantizing for maximum output entropy*, IEEE Trans. Inf. Theory IT-17 (1971), 612.
27. J. Quiñonero-Candela et al., Eds., *Dataset Shift in Machine Learning*, MIT Press, Cambridge, MA, 2009.
28. D. Rebollo-Monedero et al., *A modification of the Lloyd algorithm for* k-*anonymous quantization*, Inf. Sci. 222 (2013), 185–202.
29. M. Rosenblatt, *Remarks on a multivariate transformation*, Ann. Math. Stat. 23 (1952), 470–472.
30. S. Ruggieri, *Using t-closeness anonymity to control for non-discrimination*, Trans. Data Privacy 7 (2014), 99–129.
31. P. Samarati, *Protecting respondents' identities in microdata release*, IEEE Trans. Knowl. Data Eng. 13(6) (2001), 1010–1027.
32. J. Shao, *Mathematical Statistics*, 2nd ed., Springer, New York, NY, 2003.
33. M. Sugiyama, M. Krauledat, and K.-R. Müller, *Covariate shift adaptation by importance weighted cross validation*, J. Mach. Learn. Res. 8 (2007), 985–1005.
34. M. J. Swain and D. H. Ballard, *Color indexing*, Int. J. Comput. Vis. 7 (1991), 11–32.
35. L. Sweeney, *k-anonymity: A model for protecting privacy*, Int. J. Uncertain. Fuzz. 10 (2002), 557–570.
36. *U.S. Department* of *Health* and *Human Services*. Enrollment in the health insurance marketplace totals over 8 million people, United States Department of Health and Human Services, Press Release, 2014.
37. D. Wei et al., *Multiplicative regression via constrained least squares*, in *Proceedings of the IEEE Workshop on Statistical Signal Processing (Gold Coast, Australia, 2014)*, IEEE, 2014, 304–307.
38. D. Wei, K. N. Ramamurthy, and K. R. Varshney, Health insurance market risk assessment: Covariate shift and k-anonymity, in *Proceedings of the SIAM International Conference on Data Mining (Vancouver, Canada, 2015)*, SIAM, 2015.

39. J. Yi, J. Wang, and R. Jin, Privacy and regression model preserved learning, in *Proceedings of the AAAI Conference on Artificial Intelligence (Québec City, Canada, 2014)*, AAAI, 2014, 1341–1347.

How to cite this article: Wei D, Natesan Ramamurthy K, Varshney KR. Distribution-preserving k-anonymity. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 2018;1–18. https://doi.org/10.1002/sam.11374

APPENDIX

APPENDIX A: PROOF OF THEOREM 2

It suffices to prove that if $\tilde{x} \in \mathscr{C}(i_1^*, \ldots, i_d^*)$, then Rosenblatt's transformation yields $\hat{x} = v(i_1^*, \ldots, i_d^*)$ with probability 1 for any $(i_1^*, \ldots, i_d^*)$ so that the cell probabilities in Equation (10) translate directly into the desired resampling probabilities. This can be done by induction over the dimensions $j = 1, \ldots, d$.

For $j = 1$, we consider the CDF $F_{\tilde{X}_1}(\tilde{x}_1)$ used in the first step of transformation (Equation (6)). For any $i_1^* = 1, \ldots, n_1$, the law of total probability gives

$$\Pr(\tilde{X}_1 \in \mathcal{I}_1(i_1^*)) = \sum_{\ell=1}^{c} p_L(\ell) \times \sum_{i_1,\ldots,i_d} \Pr(\tilde{X} \in \mathscr{C}(i_1, \ldots, i_d) \mid \ell) \times \Pr(\tilde{X}_1 \in \mathcal{I}_1(i_1^*) \mid \tilde{X} \in \mathscr{C}(i_1, \ldots, i_d)).$$

Substituting Equation (10) and using the fact that

$$\Pr(\tilde{X}_1 \in \mathcal{I}_1(i_1^*) \mid \tilde{X} \in \mathscr{C}(i_1, \ldots, i_d)) = \begin{cases} 1, & i_1 = i_1^*, \\ 0, & i_1 \neq i_1^*, \end{cases}$$

we obtain

$$\Pr(\tilde{X}_1 \in \mathcal{I}_1(i_1^*)) = \sum_{\ell=1}^{c} \frac{n_\ell}{n} \sum_{i_2,\ldots,i_d} \frac{n_\ell(i_1^*, i_2, \ldots, i_d)}{n_\ell} = \sum_{i_2,\ldots,i_d} \frac{n(i_1^*, i_2, \ldots, i_d)}{n} = p_{X_1}(v_1(i_1^*)) \tag{A1}$$

by the definition of empirical distribution. It follows that for $\tilde{x}_1 \in \mathcal{I}_1(i_1^*)$,

$$\sum_{i_1=1}^{i_1^*-1} p_{X_1}(v_1(i_1)) \leq F_{\tilde{X}_1}(\tilde{x}_1) \leq \sum_{i_1=1}^{i_1^*} p_{X_1}(v_1(i_1)),$$
$$F_{X_1}(v_1(i_1^* - 1)) \leq F_{\tilde{X}_1}(\tilde{x}_1) \leq F_{X_1}(v_1(i_1^*)). \tag{A2}$$

Combining Equation (A2) with Equations (6) and (9) for $j = 1$, we conclude that $\hat{x}_1 = v_1(i_1^*)$ with probability 1.

For $j > 1$, we consider the conditional distribution of $\tilde{X}_j$ given $\tilde{X}^{j-1} = \tilde{x}^{j-1}$. Let f denote a generic PDF. Starting from

$$\Pr(\tilde{X}_j \in \mathcal{I}_j(i_j^*) \mid \tilde{X}^{j-1} = \tilde{x}^{j-1}) = \frac{\int_{\mathcal{I}_j(i_j^*)} f_{\tilde{X}^j}(\tilde{x}^j) d\tilde{x}_j}{f_{\tilde{X}^{j-1}}(\tilde{x}^{j-1})},$$

we apply a similar total probability decomposition as above to both numerator and denominator to obtain

$$\Pr(\tilde{X}_j \in \mathcal{I}_j(i_j^*) \mid \tilde{X}^{j-1} = \tilde{x}^{j-1}) = \frac{\sum_{i_1,\ldots,i_d} \frac{n(i_1,\ldots,i_d)}{n} \int_{\mathcal{I}_j(i_j^*)} f_{\tilde{X}^j|C}(\tilde{x}^j \mid C(i_1, \ldots, i_d)) d\tilde{x}_j}{\sum_{i_1,\ldots,i_d} \frac{n(i_1,\ldots,i_d)}{n} f_{\tilde{X}^{j-1}|C}(\tilde{x}^{j-1} \mid C(i_1, \ldots, i_d))}. \tag{A3}$$

Since $\tilde{x}_{j'} \in \mathcal{I}_{j'}(i_{j'}^*)$ for $j' < j$, the PDFs in the numerator and denominator of Equation (A3) are zero unless $i_{j'} = i_{j'}^*, j' < j$, while in the numerator, the integral is also zero unless $i_j = i_j^*$, in which case $\tilde{x}_j$ is marginalized out. Hence the numerator becomes

$$\sum_{i_{j+1},\ldots,i_d} \frac{n(i_1^*, \ldots, i_j^*, i_{j+1}, \ldots, i_d)}{n} f_{\tilde{X}^{j-1}|C} \times (\tilde{x}^{j-1} \mid C(i_1^*, \ldots, i_j^*, i_{j+1}, \ldots, i_d)),$$

while the denominator is similar except for an additional sum over i_j. We now exploit the fact that $\tilde{X}^{j-1}$ is uniformly distributed conditioned on a cell, that is, with $|\mathcal{I}|$ denoting the width of interval $\mathcal{I}$,

$$f_{\tilde{X}^{j-1}|C}(\tilde{x}^{j-1} \mid C(i_1^*, \ldots, i_{j-1}^*, i_j, \ldots, i_d)) = \left(\prod_{j'=1}^{j-1} |\mathcal{I}_{j'}(i_{j'}^*)| \right)^{-1}, \quad \forall \; i_j, \ldots, i_d.$$

This implies that all remaining PDFs in the numerator and denominator have the same value, reducing Equation (A3) to

$$\Pr(\tilde{X}_j \in \mathcal{I}_j(i_j^*) \mid \tilde{X}^{j-1} = \tilde{x}^{j-1}) = \frac{\sum_{i_{j+1},\ldots,i_d} \frac{n(i_1^*,\ldots,i_j^*,i_{j+1},\ldots,i_d)}{n}}{\sum_{i_j,\ldots,i_d} \frac{n(i_1^*,\ldots,i_{j-1}^*,i_j,\ldots,i_d)}{n}} = \frac{p_{X^j}(v^j(i_1^*, \ldots, i_j^*))}{p_{X^{j-1}}(v^{j-1}(i_1^*, \ldots, i_{j-1}^*))} = p_{X_j|X^{j-1}}(v_j(i_j^*) \mid v^{j-1}(i_1^*, \ldots, i_{j-1}^*)). \tag{A4}$$

Similar to Equations (A1) and (A2), Equation (A4) implies that for $\tilde{x}_{j'} \in \mathcal{I}_{j'}(i_{j'}^*), j' = 1, \ldots, j$,

$$F_{X_j|X^{j-1}}(v_j(i_j^* - 1) \mid v^{j-1}(i_1^*, \ldots, i_{j-1}^*)) \leq F_{\tilde{X}_j|\tilde{X}^{j-1}}(\tilde{x}_j \mid \tilde{x}^{j-1}) \leq F_{X_j|X^{j-1}}(v_j(i_j^*) \mid v^{j-1}(i_1^*, \ldots, i_{j-1}^*)).$$

We combine these last inequalities with Equations (6) and (9), where $\hat{x}^{j-1} = v^{j-1}(i_1^*, \ldots, i_{j-1}^*)$ by the inductive assumption. This establishes that $\hat{x}_j = v_j(i_j^*)$ with probability 1.