Protecting Personal Medical Costs Information Using Bayesian Data Synthesis

Abstract

In this paper, we use Bayesian synthesis models to create synthetic data in attempt of protecting confidential data of contractors in the Personal Medical Costs dataset (Kaggle, 2017). While preventing personal information from being leaked to the public, we also aim to maintain the utility of the dataset for statisticians to make prediction or inference regarding associations between the contractors' spending on their medical insurance and other factors. To that end, we partially synthesize our dataset with respect to contractors' spending on medical insurance, gender and age using the Bayesian multiple linear regression, the Bayesian logistic regression and the Bayesian multinomial logistic regression models, respectively. Afterwards, we evaluate the utility and risk of the synthetic dataset and perform differential privacy for added protection to the summary statistics of the confidential dataset. After our evaluations, we conclude that we successfully created a synthetic dataset that provides better privacy protection than the confidential dataset does.

1 Introduction

Medical datasets contain a plethora of information about patients that are considered confidential, but they are also regarded as crucial sources for studying diseases, economics, and other areas of discussion. We found the Personal Medical Cost dataset from Kaggle (Kaggle, 2017) and deemed it important in terms of predicting how much an individual spends on medical insurance, and other inferences. We hope to provide this dataset with protection so that when the dataset is released to the public, the contractors' private personal information will not be leaked yet users can examine associations between medical cost and contractors' attributes.

To that end, we employed Bayesian statistical models to partially synthesize information about variables that we deemed sensitive. Afterwards, we conducted utility and risk evaluations to examine whether our synthetic dataset could serve for analytical purposes while also providing sufficient protection of the confidential information. We also applied differential privacy methods to further protect the summary statistics of the confidential dataset. In the following sections of this paper, these steps will be elaborated.

1.1 The Dataset

The Personal Medical Cost dataset has 1338 participants and 7 variables: `age`, `sex`, `bmi`, `children`, `smoker`, `region`, and `charges`. Here, `age` is the age of the primary beneficiary, `sex` is the gender of insurance contractor, `bmi` is the body mass index, `children` is the number of children covered by health insurance, `smoker` suggests the smoking status of insurance contractor, `region` is the beneficiary's residential area in the United States, and `charges` is individual medical costs billed by health insurance (Kaggle, 2017). In our research, all variables are considered. `sex` and `smoker` are binary variables, `region` is categorical, `age` and `children` are ordinal, and the remaining variables are continuous. Table 1 presents the details of the variables used, including variable type and description.

Variable	Type	Description
Charges	Continuous	Range from 1,122 to 63,770
BMI	Continuous	Range from 15.96 to 53.13
Smoker (yes)	Binary	yes/no
Region	Categorical	northwest, northeast, southwest, southeast
Age	Ordinal	Range from 18 to 64
Children	Ordinal	Range from 0 to 5
Sex	Binary	1=male, 2=female

Table 1: Table detailing the data

Generally, people tend to keep their medical information private, so we deemed insurance charges to be the most sensitive variable. We assumed that if an intruder knew a participant's age and sex - and maybe other personal information outside of this dataset - they may be able to identify the person and hence their spending on medical insurance. Therefore, we also considered `age` and `sex` as sensitive variables.

Before the synthesis models were built, some data processing had to be completed. As a continuous variable, `charges` did not appear to be normally distributed, as shown in Figure 1. It was noted that `charges` was significantly skewed, and log transformation was insufficient to normalize the data. Therefore, we obtained a newly created variable called `transformed_charges` by transforming the `charges` data using the bestNormalize package (Peterson, 2022), which is expected to normalize data using the optimal method selected from a set of transformation-estimating functions. As shown in Figure 2, applying the bestNormalize function to our dataset successfully made transformed_charges normally distributed.

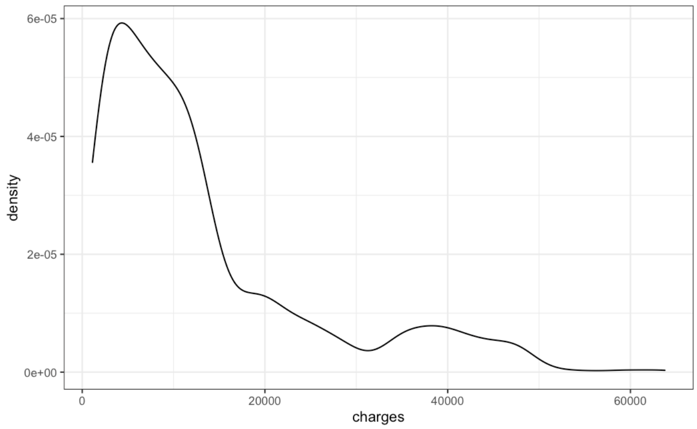

Figure 1: Distribution of charges

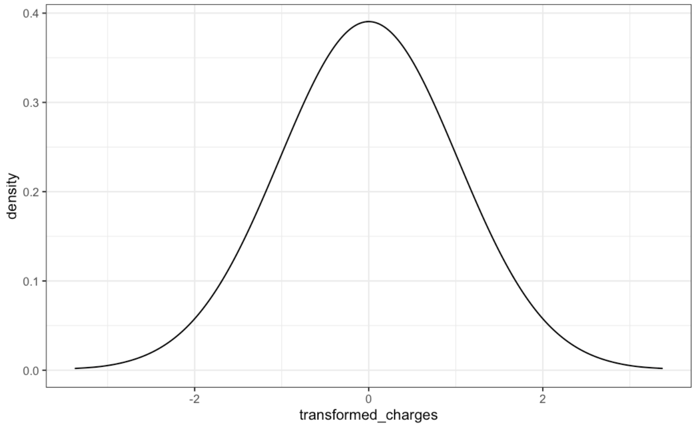

Figure 2: Distribution of transformed_charges

We also made age a categorical variable, age_cat; otherwise, an ordinal varaible with enormous levels would be difficult as well as unreasonable to deal with. We experimented with several different category widths and number of categories and found that 3 categories - younger than 30, between 30 and 50, and older than 50 - yielded the optimal utility results, which will be demonstrated in the later section of this paper. Lastly, sex, as a binary variable, did not require transformation to be fit into a model.

1.2 The Synthetic Data Approach

To enhance confidentiality of the medical cost dataset, we applied the partial synthesis approach to our dataset and sequentially synthesized data regarding insurance charges as well as age and sex of the insurance contractors. By the partial synthesis approach, a set of sensitive variables and key identifiers are synthesized, while other attributes remain unchanged (Little, 1993). To generate the synthetic dataset, we at first fit Bayesian models on the confidential data to estimate the posterior distributions. Given the posterior predictive distributions, we then simulated synthetic values for the sensitive variables. With careful selection and design of statistical models, this method preserves crucial statistical features of the confidential dataset, such as mean, variance and joint probability distribution and further allows for making inference and prediction. Meanwhile, it reduces risks of identification and attributes of subjects being predicted by intruders (Drechsler, 2011).

Given that transformed_charges is continuous, sex is binary and age_cat is categorical, we adopted the Bayesian multiple linear regression, the Bayesian logistic regression and the Bayesian multinomial regression models, respectively. To fit and run our models, we used the R package brms, which provides an interface to fit Bayesian generalized (non-)linear multivariate multilevel models using Stan (Bürkner, 2022). In the remaining part of our paper, we will elaborate the model implementations in Section 2, perform utility and risk evaluation in Section 3 and 4, apply differential privacy to our confidential data in Section 5 and eventually, make closing remarks in Section 5.

2 The Synthesis Model

We applied different Bayesian models to synthesize transformed_charges, sex, and age_cat in a sequential manner. Although transformed_charges is regarded as the most sensitive variable, it was synthesized first; because the other two variables are either categorical or binary, we assumed their confidential and synthetic data would be less similar to each other than that of transformed_charges. Thus, if we synthesized transformed_charges last using the synthesized sex and age_cat, while the transformed_charges data would be better protected, we assumed utility of our dataset would be worse-off as the synthesized transformed_charges data of individual records could be significantly deviated from the confidential ones. Because we still hope to maintain utility of our dataset, we decided to synthesize the sensitive variables in the order stated earlier.

Given that transformed_charges is continuous and normally distributed, we deemed that the Bayesian multiple linear regression would be the most appropriate method for the data synthesis purposes. The model we employed to synthesize transformed_charges is stated below:

$$\texttt{transformed\_charges} \sim \texttt{1 + bmi + smoker + region},$$

as we deemed that among non-sensitive variables, bmi, smoker, and region are the most relative to the specific value of the medical insurance cost. We then ran the simulation through 5000 iterations and drew the last 2000 to generate our synthetic data (Note that number of iterations run and number of iterations used as warm-up were adjusted according to model performance, which was reflected from the trace and autocorrelation diagnostic plots). Afterwards, we compared density plots of synthesized and confidential transformed_charges as shown in Figure 3. The comparison suggests that our synthesis result is promising, as the synthesized transformed_charges successfully captures the general distribution of our confidential transformed_charges data.

We further denormalized transformed_charges back to the normal scale of medical insurance costs using the bestNormalize package in order to examine similarity between density plots of synthesized and confidential charges data as shown in Figure 4. The comparison suggests that our synthesis result as well as the normal transformation method are successful. Despite peak on the left of the confidential charges distribution, the synthetic result captures the general confidential charges distribution.

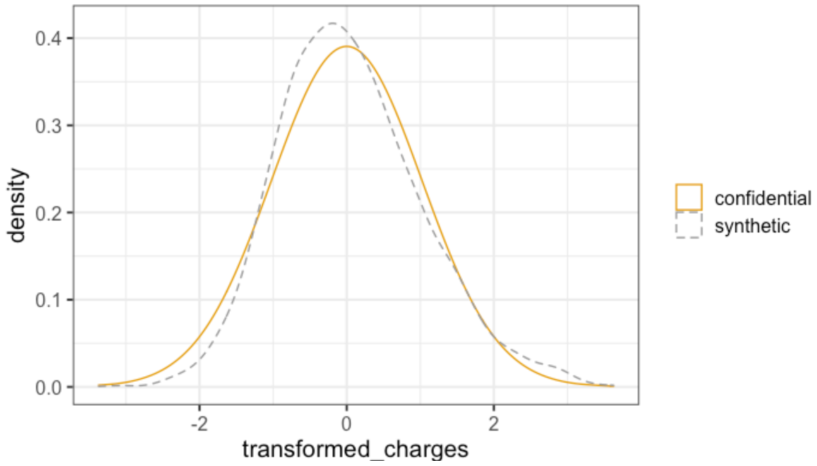

Figure 3: Utility Check for Synthesized `transformed_charges`

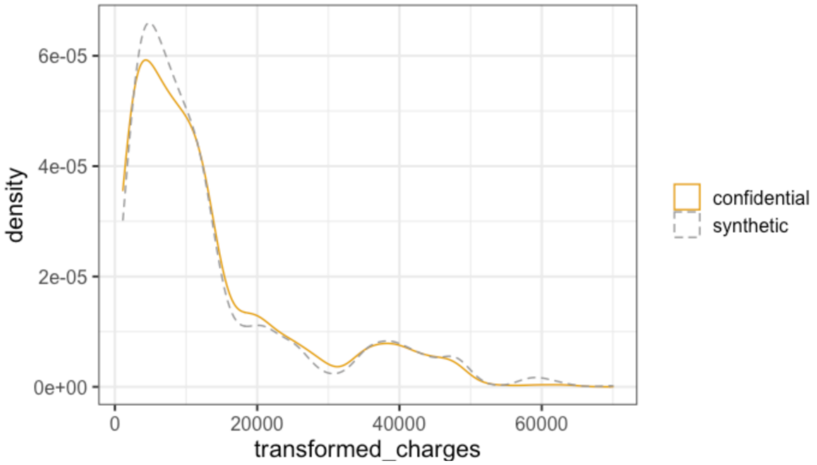

Figure 4: Utility Check for Synthesized `charges`

Then, we further synthesized the second sensitive variable, `sex`. Because `sex` is a binary variable, we considered the Bayesian logistic regression as the most appropriate method for the data synthesis purposes. The model we employed to synthesize `sex` is stated below:

$$\texttt{(sex - 1)} \sim \texttt{1 + bmi + smoker + region + syn\_t\_charges},$$

where `syn_t_charges` is the synthesized `transformed_charges`. We then ran the simulation through 5000 iterations and drew the last 2000 to generate our synthetic data. Afterwards, we compared density plots of synthesized and confidential `sex` as shown in Figure 5. The comparison suggests that our synthesis result is still promising, as counts of either gender category between confidential and synthetic `sex` are fairly close to each other, and thus, the synthetic data successfully captures general distribution of the confidential data.

Lastly, we synthesized the third sensitive variable, `age_cat`. As `age_cat` is a categorical variable, we employed the Bayesian multinomial logistic regression model to synthesize our data. The model we employed to synthesize `age_cat` is stated below:

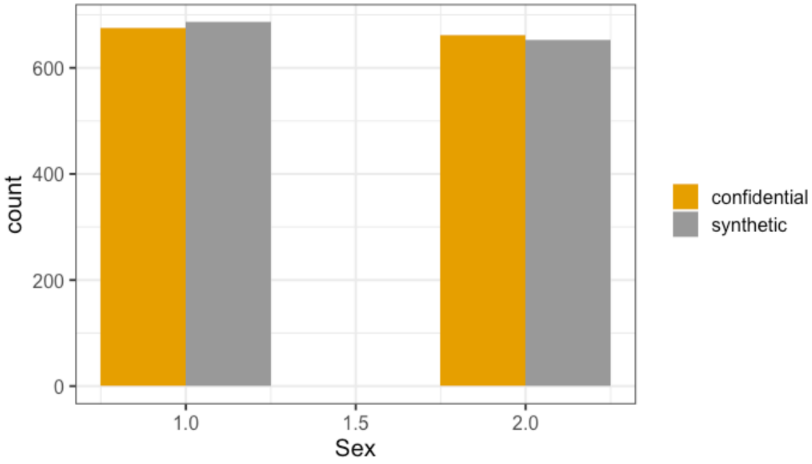

Figure 5: Utility Check for Synthesized `sex`

$$\texttt{age\_cat} \sim \texttt{1 + bmi + smoker + region + syn\_t\_charges + syn\_sex},$$

where `syn_t_charges` is the synthesized `transformed_charges`, and `syn_sex` is the synthesized `sex`. We then ran the simulation through 5000 iterations and drew the last 2000 to generate our synthetic data. Afterwards, `age_cat`, we compared density plots of synthesized and confidential `age_cat` as shown in Figure 6. The comparison, however, suggests that our synthesis model performed poorly. The density plots show that within each age category, counts of the synthetic and confidential data are vastly different from each other. We deemed that it could be explained by the fact that `age_cat` was synthesized using the synthesized `transformed_charges` and `sex` data, as minor differences between the confidential and synthetic data could accumulate and eventually cause such significant gap.

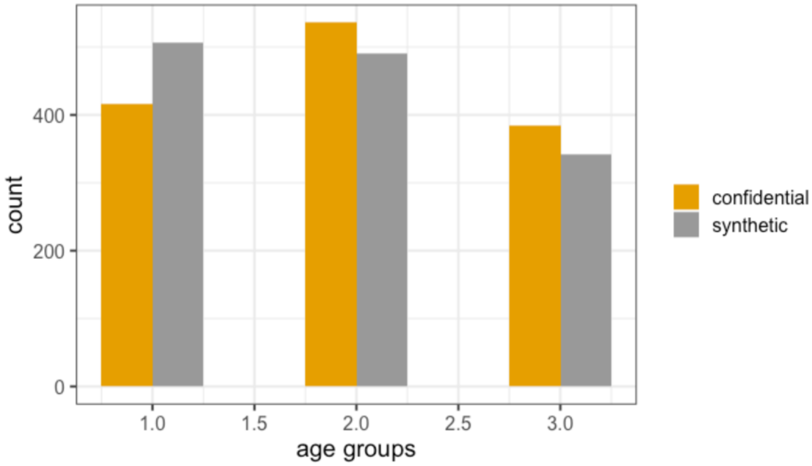

Figure 6: Utility Check for Synthesized `age_cat`

3 Utility Evaluation

After producing the synthetic Personal Medical Cost dataset, it was imperative for us to check the utility-risk trade-off to evaluate how useful our model is, as well as how well it protects the participants' information. First, utility is evaluated. It was important to evaluate both global utility as well as analysis-specific utility. While evaluating global utility, we examine the similarity between the confidential data distribution and the synthetic data distribution. While evaluating analysis-specific utility, we examine whether researchers can obtain statistical inferences on the synthetic data that are similar to those obtained on the confidential data.

In these evaluations, we first made sure to denormalize `transformed_charges` back to the normal scale of medical costs so that our interpretations would be meaningful.

3.1 Global Utility

First focusing on global utility, we used the pMSE and eCDF, which focus on identifying the distributional differences between the confidential and the synthetic data.

3.1.1 pMSE

Propensity scores examines the probability for individuals in the dataset being part of the generated synthetic dataset given their other variables. If the probability distributions of the confidential and synthetic groups differ, then the observations in the two groups differ in their covariates (Woo et al., 2009; Snoke et al., 2018).

Based on this idea, we could compare distributions of the propensity scores in both confidential and synthetic datasets by calculating the propensity score mean-squared error (pMSE), measuring the mean-squared difference between the estimated propensity probabilities and the probability of a record being synthetic if original and synthetic observations were interchangeable. Assume we merge the confidential as well as the synthetic datasets with n_c and n_s records, respectively. Then, we introduce a new variable S. For each record i, we set $S_i = 0$ if it is from the confidential dataset and $S_i = 1$ if it is from the synthetic dataset. Afterwards, we estimate the probability of each record being in the synthetic dataset - the estimated propensity score $\hat{p}_i$ - by fitting a model using available predictors in our datasets. Eventualy, we obtain pMSE following the formula:

$$pMSE = \frac{1}{n_c + n_s} \sum_{i=1}^{n_c+n_s} (\hat{p}_i - c)^2 \text{ (Woo et al., 2009; Snoke et al., 2018)},$$

where c is the proportion of records with synthetic data.

Larger values of pMSE indicate a low level of utility, and smaller values of pMSE indicate a high level of utility. Our pMSE was calculated to be 0.01524864, which is a low value, and thus indicates that our synthetic data has high global utility.

3.1.2 eCDF

The empirical cumulative distribution function (eCDF) is the distribution function associated with the empirical measure of a sample and is a discrete distribution function which considers every observation in the sample to be an equally likely outcome. In the sense of global utility, the eCDF relies on comparing the empirical distributions of the synthetic dataset with that of the confidential dataset; similar samples should have similar eCDF's. Further, U_m is the maximum absolute difference between the empirical CDFs, and U_a is the average squared differences between the empirical CDFs. Given the merging dataset from the pMSE case, if we denote percentile under the empirical CDF distribution of each record from the confidential dataset as p_i^C and that from the synthetic dataset as p_i^S, we may obtain U_m and U_a using the formula below:

$$U_m = \max_{1 \leq i \leq (n_c+n_s)} \| p_i^C - p_i^S \|,$$

$$U_a = \frac{1}{(n_c+n_s)} \sum_{i=1}^{(n_c+n_s)} (p_i^C - p_i^S)^2 \text{ (Woo et al., 2009).}$$

Smaller values of U_m and U_a mean that there is a higher similarity level between the confidential and the synthetic data, indicating high utility, and larger values indicate lower utility. The table below displays our values of U_m and U_a for our sensitive variables.

Variable	U_m	U_a
Charges	0.04035874	0.0002792408
Sex	0.007473842	2.843013e-05
Age	0.06651719	0.001922093

Table 2: U_m and U_a values in eCDF global utility evaluation

The table shows that our values of U_m and U_a are small and close to 0, indicating that our synthetic dataset has high global utility. This agrees with the results of our pMSE test and thus, since both methods indicate high utility, we can conclude that our dataset has high global utility.

3.2 Analysis-Specific Utility

On an analysis-specific level, it is important to understand what analysis a user might perform on the data, and check the utility of the synthetic data in this context. Assume that the user wants to study the mean and 95% confidence intervals of medical insurance costs and make inference or prediction about it using all other available variables in the dataset. We evaluate the analysis-specific utility of our synthetic data using these measures as well as interval overlaps.

3.2.1 Inference on Mean Estimates

First, we found the mean and 95% confidence intervals of `charges` for our synthetic dataset, and then did the same for our confidential data. We then compared these values as shown in the table below. The mean and 95% confidence intervals for charges are similar between the synthetic dataset and confidential dataset, indicating high utility.

Variable	Mean	2.5%	97.5%
Synthetic	$13,599.66	$12,908.45	$14,290.87
Confidential	$13,270.42	$12,620.95	$13,919.89

Table 3: Inference for Average Charges

3.2.2 Inference for Regression Coefficients

Next, we examined the analysis-specific utility of our synthetic dataset based on regression coefficients. It was assumed that the analyst would perform exploratory data analysis before running regression and find charges significantly skewed, as shown in Figure 1, and thus use a log transformation on the confidential `charges` data. Therefore, we used our synthetic data with a log transformation on the denormalized `charges` and ran a regression with `charges` as the Y output and all the other variables as covariates, and compared the resulting coefficients with those obtained from the same regression on the confidential data.

	Coefficient	2.5%	97.5%
BMI	0.0251	0.0195	0.0306
Smoker(yes)	1.4388	1.3576	1.5200
Northwest	-0.0201	-0.1131	0.0728
Southeast	-0.2171	-0.3110	-0.1232
Southwest	-0.1411	-0.2350	-0.0473
Age_cat2	0.3550	0.2770	0.4329
Age_cat3	0.7107	0.6229	0.7985
Sex	0.1413	0.0753	0.2073

Table 4: Synthetic Regression Coefficients

	Coefficient	2.5%	97.5%
BMI	0.0139	0.0094	0.0184
Smoker(yes)	1.5546	1.4891	1.6201
Northwest	-0.0540	-0.1294	0.0214
Southeast	-0.1614	-0.2372	-0.0855
Southwest	-0.1211	-0.1968	-0.0454
Age_cat2	0.6781	0.6153	0.7409
Age_cat3	1.1874	1.1190	1.2558
Sex	0.0740	0.0212	0.1267

Table 5: Confidential Regression Coefficients

After comparing the coefficients and 95% confidence intervals of the predictors, it was found that the coefficients of many variables were very similar, which indicates high utility. However, `age_cat2`, `age_cat3`, and `sex` from regressions run on the confidential and the synthetic data had very different coefficients; the synthetic coefficients for `age_cat2` and `age_cat3` were double that of the confidential values, and vice versa for `sex`. These differences could be because `age_cat` and `sex` were synthesized after `transformed_charges` and their models were built based on confidential and synthetic data. Specifically, `age_cat` was synthesized last so it was built upon synthesized `transformed_charges` and synthesized `sex`. Thus, their models were not as good as they would have been if they were synthesized first. From running these regressions, it is indicated that our synthetic data has somewhat higher utility in running regression for the medical insurance costs.

3.2.3 Interval Overlap Utility Measure

Lastly, we evaluated interval overlap to calculate and compare the closeness between intervals. There are two definitions of this; by definition 1, no overlap returns 0, but by definition 2, as the intervals get farther apart, the interval overlap returns a more negative number. In order to understand the magnitude of the low utility if we had no overlap, it was decided to use definition 2.

Let (L_S, U_S) denote the $(1-2\alpha)\%$ confidence interval for the estimand from m synthetic data, $Z = (Z^1, \cdots, Z^m)$ and (L_C, U_C) denote that from the confidential data. According to definition 2, the interval overlap measure is obtained by

$$IO = \frac{1}{2}\left(\frac{\min(U_C, U_S) - \max(L_C, L_S)}{U_C - L_C} + \frac{\min(U_C, U_S) - \max(L_C, L_S)}{U_S - L_S}\right)$$

(Snoke et al., 2018; Ros, Olsson, Hu, 2020).

Variable	Interval Overlap
Charges	0.7551551
BMI	-0.0999901
Smoker (yes)	0.2130744
Northwest	0.8077085
Southeast	0.679256
Southwest	0.8921794
Age_cat2	-1.311105
Age_cat3	-2.084004
Sex	0.4382989

Table 6: Interval Overlap - Definition 2

The results in Table 6 show that `charges` and all levels of `region` have high overlap, but some other variables have very small positive values or negative values. Specifically, both levels of `age_cat` have large negative overlap, This could also be due to `age_cat` being the last variable to be synthesized, and thus the model is not as good.

While the mean inference indicated high utility, the inference for regression coefficients showed somewhat high utility and the interval overlaps showed mediocre utility. Overall, our global and analysis-specific utility evaluations indicated that our synthetic data has somewhat high utility. To understand our utility-risk trade-off, risk also needs to be evaluated.

4 Risk Evaluation

In this section, we will evaluate two types of disclosure risks on our synthetic dataset: identification and attribute disclosures. For identification disclosure risk evaluation, we examine how well the intruder could correctly identify records of interest in the synthetic data. For attribute disclosure risk evaluation, we examine who well the intruder could correctly infer the true confidential values of the synthetic records given information from the synthetic data.

4.1 Identification Disclosure Risk

Identification disclosure risk is usually evaluated using the matching-based approaches. Such methods, at first, make assumptions regarding information that the intruder knows for a confidential record. Then, we examine which records in the synthetic data are matched with that record based on our assumptions. Eventually, we look into whether the matched records contain true match and how unique the true match is. To reach that goal, we can either investigate the three risk summaries - the expected match risk, the true match rate, and the false match rate - or apply the record linkage approaches. In our case, we assume the intruder possesses knowledge about smoking status, living region, and number of children of an arbitrary record in the confidential data, as these information can be easily acquired.

4.1.1 Three Risk Summaries

We will begin with examining the three measures of identification disclosure risk. At first, the expected match risk examines the mean likelihood of the correct match to be found for each record, and for the sample as a whole. Then, the true match rate measures the proportion of existing unique matches. Furthermore, the false match rate measures the proportion of false matches within unique matches (Reiter, 2005). To sum up, identification disclosure risk for the sample is higher given higher expected match risk, higher true match rate, and lower false match rate.

A drawback of this method is that the identification risk calculation algorithm performs poorly if the target variables contain both continuous and categorical variables. Therefore, we examine them separately, assuming known variables are `smoker`, `region`, and `children`. While focusing on categorical variables `sex` and `age_cat`, we obtained an expected match risk of 32.27, a true match rate of 0.2990%, and a false match rate of 92.73% for the synthetic dataset, and an expected match risk of 208, a true match rate of 3.438%, and a false match rate of 0% for the confidential dataset.

While focusing on the continuous variable `charges`, we used a radius of 0.2 to declare a match. we obtained an expected match risk of 39.96, a true match rate of 0.5232%, and a false match rate of 85.11% for the synthetic dataset, and an expected match risk of 187.63, a true match rate of 2.317%, and a false match rate of 0% for the confidential dataset. In either case, differences between risk summaries derived from the synthetic and confidential datasets indicate that the identification disclosure risk in our synthetic data is low.

4.1.2 Record Linkage Approaches

William E. Winkler (2004) introduced record linkage approaches as metrics of identification disclosure risk, which can be applied to linking records in the synthetic data to those in the confidential data. Among all linkages, identification disclosure risk can be examined in terms of true and false links, and high percentage of true links or low percentage of false links suggests high identification disclosure risk.

To apply this method to our datasets, we employed the `reclin2` package (Van Der Lann, 2022). At first, given the assumptions of the intruder, we generated pairs between the known and synthesized variables. Then, we compared values of the synthesized variables and calculated similarity score of each pair. Given the weight of each pair, we further selected one-to-one linkages between variables available to the intruder from external databases and those from the synthetic dataset and eventually calculated the percentages of true links and false links (Winkler, 2004).

While applying the procedures to our synthetic dataset, we obtained a true linkage rate of 4.0359% and a false linkage rate of 95.9641%. However, applying the same procedures to the confidential dataset yielded a 100% true linkage rate. Therefore, the result yielded from the linkage record approaches is consistent with that from the three risk summaries, and thus, our synthetic data provides better identification protection compared to the confidential data.

4.2 Attribute Disclosure Risk

Attribute disclosure risk is the risk of inferring the confidential values of the synthesized variables in the synthetic data by the intruder. It is assumed that an intruder knows some features of the target individual, and these features are called key variables. It is further assumed that the intruder will infer other features based on their information about the key variables, and these features are called target variables. The attribute disclosure risk is usually evaluated using the CAP risk measure and the classification-based risk measure.

4.2.1 Average CAP

The correct attribution probability (CAP) measure was introduced by Elliot (2014) and Taub et al. (2018). It examines the probability of correctly predicting the value of the target variable of a subject using the empirical distribution of the variable among synthetic observations with the same key variables. Furthermore, the average CAP takes the mean of all individuals' CAPs, and it can be interpreted as the fraction of valid predictions for an intruder who is making prediction of the target variable from each record in the confidential dataset by sampling from the empirical distribution of the target variable within the synthetic data, holding the value of the key variable constant (Elliot, 2014; Taub et al., 2018).

In our case, we assumed the intruder has information about the subject's smoking status, living region, and number of children, which are the key variables. Then, we calculated the average CAPs for `charges`, `sex`, and `age_cat`, the target variables, using both the confidential as well as the synthetic data. While applying the calculation to our synthetic dataset, the average CAPs for our target variables are 0.1862, 0.5045, and 0.3380, respectively. While applying the calculation to our confidential dataset, the average CAPs for our target variables are 0.2422, 0.5172, and 0.3937. Through comparison, we found that although differences are slight, in general our synthetic dataset could provide better attribute protection compared to the confidential data.

4.2.2 Classification-Based Risk Measure

The CAP statistic has a weakness that it predicts the values of the target variable using simple model. To address this, Choi et al. (2017) and Kaur et al. (2021) proposed to evaluate the attribute disclosure risk using more general classification approaches. This method, at first, trains a classifier that predicts the values of the target variable given the values of keys using the synthetic dataset. Then, the classifier predicts the values of the target variable using the confidential data. Finally, the average attribute disclosure risk is calculated as the fraction of records for which the prediction is valid. In our case, we used a k-NN classifier considering 10 nearest neighbors for each record to predict `charges`, `sex`, and `age_cat`, separately.

At first, we compared the mean-square error (MSE) and the predictive error rate of the predictions using the k-NN classifiers trained on the synthetic and the confidential datasets. Predictions on `charges` made by the model trained on the synthetic data yielded a MSE of 53129895, and those made by the model trained on the confidential data yielded a MSE of 32072018. Predictions on `sex` made by the model trained on the synthetic data yielded an error rate of 0.4836, and those made by the model trained on the confidential data yielded an error rate of 0.3789. Predictions on `age_cat` made by the model trained on the synthetic data yielded an error rate of 0.5874, and those made by the model trained on the confidential data yielded an error rate of 0.0.3595. In general, the accuracy of predicting the target variables is lower while using the synthetic dataset to train the classifier, which demonstrates that our synthetic data provides better protection compared to the confidential data. The same conclusion can be made while examining the fraction of records that have a less accurate prediction with the synthetic data compared to with the confidential data, and they are 0.3356, 0.4439, 0.4776 for the models predicting `charges`, `sex`, and `age_cat`, respectively.

As a result, our synthetic data provides better identification and attribute protection compared to the confidential data, and thus has low risk. In the next section, we will move forward with providing the summary statistics of the confidential dataset with protection by adding random noise.

5 Differential Privacy

Differential Privacy, a formal mathematical framework to provide privacy protection guarantees, was also calculated. The main idea of differential privacy is to add random noise to the output of summary statistics from confidential data that would be useful for users to know. Specifically, with given privacy budget - a term ε that is to be spent by the database holder when calculating statistics - we can add noise in order to preserve privacy (Dwork, McSherry, Smith, 2006). When noise increases, sensitivity increases and the privacy budget decreases. There are three types of differential privacy: sequential composition, parallel composition, and post-processing.

Sequential composition is differential privacy that is performed on overlapping subsets, such as continuous variables. Hence, we used this to calculate the differential privacy statistics of the mean and median of `charges`. For the mean, we used a privacy budget of 0.1. However, for the median we had to use a much larger privacy budget of 1000 because the calculation had a much wider range than the mean, but we wanted a similar amount of noise.

Statistic	Epsilon	With Noise	Without Noise
Mean	0.1	$13,593.20	$13,270.42
Median	1000	$9,116.022	$9,382.033

Table 7: Sequential composition for average charges

With an epsilon of 0.1, the mean with and without noise is similar enough to be useful but not so similar that it is the confidential data, and likewise for the mean with an epsilon of 1000.

Parallel composition is another method of differential privacy but conversely to sequential, it is used for non-overlapping subsets. Hence, it was used to calculate average `charges` per category of `age_cat` and `sex`, again with ε of 0.1.

Category	Epsilon	With Noise	Without Noise
Male	0.1	$14,118.14	$13,956.75
Female	0.1	$11,575.52	$12,569.58

Table 8: Average charges for male and female

Even though the differential statistic for average `charges` for people older than 50 wasn't as similar to the true value, generally, the results are similar to our true confidential data and make logical sense: as

Category	Epsilon	With Noise	Without Noise
Younger than 30	0.1	$9,343.874	$9,182.487
30-50	0.1	$12,129.53	$13,123.59
Older than 50	0.1	$13,825.43	$17,902.55

Table 9: Average charges per age category

you get older, your medical charges would typically increase. Women having lower medical charges makes some sense as one could hypothesize that women are less likely than men to go to the doctor for smaller issues and, typically, their behavior is less risky so they don't need to see the doctor as often.

Lastly, post-processing was used to add noise to the counts of categorical/binary variables, such as `age_cat` and `sex`, as we assumed a user may also want to use this key data in their analysis. The count with noise is, again, similar enough to the count without noise to be useful, but not so similar that it is the same as the true value.

Category	Epsilon	With Noise	Without Noise
Younger than 30	0.1	417	422
30-50	0.1	536	660
Older than 50	0.1	385	236
Male	0.1	676	678
Female	0.1	662	660

Table 10: Post-processing counts for age and sex

6 Conclusion

Personal medical charges is typically information that people want to be kept private, and for this study we attempted to accomplish this data protection; we created a synthetic dataset that balances utility and risk by providing useful data without risking any participants' identity. We also calculated differential privacy statistics to protect partipicants' identities while also providing useful information to analysts. After evaluating all of these results, we conclude that the synthetic data provides sufficient privacy protection for participants of the Personal Medical Costs dataset.

References

Bürkner, P. C. 2022. ”Bayesian Regression Models using Stan”. Embracing Uncertainty. https://paul-buerkner.github.io/brms/

Choi, M. 2018. ”Medical Cost Personal Datasets”. Kaggle. https://www.kaggle.com/datasets/mirichoi0218/insurance

Drechsler, J. (2011). Synthetic datasets for Statistical Disclosure Control: Theory and implementation. Springer.

Dwork, C., F. McSherry, K. Nissim, and A. Smith. 2006. “Calibrating Noise to Sensitivity in Private Data Analysis.” Proceedings of the Third Conference on Theory of Cryptography, 265–84.

Elliot, M. 2014. “Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team.” CMIST.

Laan, J. van der. 2022. ”Record linkage toolkit [R package reclin2 version 0.2.0]”. The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/reclin2/index.html

Little, R. J. A. 1993. “Statistical Analysis of Masked Data.” Journal of Official Statistics 9: 407–26.

Peterson, R. A. 2022. ”Using the bestnormalize package”. https://cran.r-project.org/web/packages/bestNormalize/vignettes/bestNormalize.html

Reiter, J. P. 2005. “Estimating Risks of Identification Disclosure in Microdata.” Journal of the American Statistical Association 100: 1103–12.

Ros, K., H. Olsson, and J. Hu. 2020. “Two-Phase Data Synthesis for Income: An Application to the NHIS.” Privacy in Statistical Databases (e-Proceedings).

Snoke, J., G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic. 2018. “General and Specific Utility Measures for Synthetic Data.” Journal of the Royal Statistical Society, Series A (Statistics in Society) 181: 663–88.

Taub, J., M. Elliot, M. Pampaka, and D. Smith. 2018. “Differential Correct Attribution Probability for Synthetic Data: An Exploration.” Privacy in Statistical Databases, 122–37.

Winkler, William E. 2004. “Re-Identification Methods for Masked Microdata.” In Privacy for Statistical Databases, edited by J. Domingo-Ferrer and V. Torra, 216–30.

Woo, M. J., J. P. Reiter, A. Oganian, and A. F. Karr. 2009. “Global Measures of Data Utility for Microdata Masked for Disclosure Limitation.” The Journal of Privacy and Confidentiality 1: 111–24.