Bayesian Data Synthesis for Protecting Sensitive Salary Data Information

December 2022

Abstract

We propose a Bayesian data synthesis process for synthesizing a person's income, a sensitive and confidential variable. We consider four differently-sized synthetic models, found from using best subsets, and evaluate the respective utility and risk of each. To determine whether sufficient privacy protection has been achieved, we compare each synthetic model's utility and risk to that of the confidential model. We conclude that the 2-variable synthetic model has strong utility and low risk, indicating sufficient data privacy protection is in place. We present the utility and risk evaluations by synthesizing data from Glassdoor.

Keywords: data privacy, Bayesian methods, utility, disclosure risk, synthetic data

1 Introduction

There has long been an issue of releasing data which would be interesting and useful to study, but may harm the individuals whose entries comprise this data. Although deemed a relatively emergent challenge, the confidentiality issue present with individual-level data (or microdata) has been studied and explored by numerous statisticians. Little (1993), Rubin (1993), Raghunathan et al. (2003), and Drechsler (2011) overcame this microdata obstacle by simulating the values of sensitive variables using statistical models and, rather than releasing the confidential records, releasing the synthetic records. This protected the privacy of the individuals who may have been harmed with the release of the original dataset while also maintaining the useful analysis qualities of the original dataset (such as the mean and the relationships between variables).

Among the different types of data which may be sensitive, salary data is one of the most common. A person's salary is generally considered private unless that person wishes to share; within a company, salary data is kept confidential. There may be risk of physical harm if a person is upset at their making less money than a coworker, for example, and harassment that is salary-related is another possibility.

Throughout this paper, we use a Bayesian synthetic data approach to protect salary information. We find that our synthetic data has very strong utility and risk improves significantly.

1.1 Data

We found our dataset on Kaggle, a site that boasts free and easy-to-use datasets (Jauhari, 2019). Our data is titled "Glassdoor- Analyze Gender Pay Gap" and contains the variables JobTitle, Gender, Age, PerfEval, Education, Dept, Seniority, BasePay, and Bonus. Specific details about all of these variables can be found in the table below.

Table 1: Variables in the Glassdoor dataset

Variable Name	Variable information
JobTitle	Categorical; the specific job title: 1 = Data Scientist, 2 = Driver, 3 = Financial Analyst, 4 = Graphic Designer, 5 = IT, 6 = Manager, 7 = Marketing Associate, 8 = Sales Associate, 9 = Software Engineer, 10 = Warehouse Associate.
Gender	Categorical; gender: 1 = male, 2 = female.
Age	Continuous; age (ranges from 18 - 65 years).
PerfEval	Categorical; performance evaluation score (1-5): 1 indicates the poorest performance while 5 indicates the greatest performance.
Education	Categorical; the highest level of education reached: 1 = High School, 2 = College, 3 = Masters, 4 = PhD.
Dept	Categorical; the department categorization: 1 = Administration, 2 = Engineering, 3 = Management, 4 = Operations, 5 = Sales.
Seniority	Categorical; seniority level (1-5): 1 indicates the most junior while 5 indicates the most senior.
BasePay	Continuous; Base pay received in a calendar year (in USD).
Bonus	Continuous; Bonus received in a calendar year (in USD).

We deemed the BasePay variable as the one most sensitive, choosing to synthesize that. BasePay refers to someone's base salary (not accounting for, say, their bonus).

The remainder of this paper is organized as follows: In section 2, we go through the methods for our synthesis of salary data, along with the methods we use for evaluating the utility and risk of our model. In section 3, we conduct several evaluations of our model's utility-risk tradeoff and give differential privacy results. Finally, in section 4, we discuss our model's functionality bearing all of these results in mind.

2 Methods

2.1 Synthesis Model & Implementations

The first step in our synthesis process is determining the predictors in our final model. Accordingly, we begin with graphical exploration of our data. We experiment with scatterplots and box plots attempting to find potential patterns between our outcome variable, BasePay, and each predictor. Not having tremendous success, we decide to use best subsets, an automated variable selection algorithm, to choose our predictors in our final model based on adjusted r-squared. The idea behind best subsets is to fit all 2^p, where $p =$ the number of predictors in the model, possible models and find the best one. We ultimately settle on a model with two predictors (more detail surrounding this decision can be found in section 4): Age and Seniority. From here, we begin prepping for the synthesis process by creating a design matrix based on the chosen model: $BasePay \sim Age + Seniority$. We then run the Bayesian multiple linear regression on our selected predictors using default priors. From here, we save the posterior parameter draws of estimated parameters. We use these draws to generate synthetic data given the posterior predictive distribution. Next, we check the trace and autocorrelation diagnostic plots. Traceplots work by plotting the parameter values against each MCMC iteration number and seeing if the graph instantly dips towards zero. Autoccorelation plots work by checking to see if the posterior draws are relatively independent from each other. We find encouraging results from the trace plot—that is, the results are quite random and sporadic. However, the autocorrelation is not as promising—the plot didn't fall to zero as quickly as we would like. Thus, we go back and increase our thinning value in the Bayesian linear regression. With these now satisfying results, we then create our synthesis function that draws from the posterior parameter draws of the estimated parameters, and then we perform the synthesis for our dataset by using one set of posterior draws at the index iteration of the MCMC. We then conduct a Bayesian linear regression utility check by creating a density plot comparing synthetic and confidential BasePay. Rather than synthesizing just one dataset, we synthesize $m = 20$ datasets, then create a density plot displaying the first three synthetic BasePay datasets to check our utility against the confidential BasePay (Figure 1). Findings from the density plots are discussed in section 3.

2.2 Global Utility

Global utility evaluation is focused on the closeness of the synthetic data with the confidential data. These evaluations use common statistical tools, such as data modeling, in order to identify the distribution differences between the two datasets. We use two evaluation methods to measure our global utility: pMSE and eCDF.

2.2.1 pMSE

Propensity scores are used for measuring the probability that individuals were assigned to a specific group based on the values of the covariates. Differing probabilities suggest that the individuals in the two groups vary across these variables. In Woo et al. (2009) and Snoke et al. (2018), propensity score matching is suggested to see if the confidential and synthetic values differ from one another. Here, the assigned group is whether the observation appears to be part of the generated synthetic dataset.

To calculate pMSE, we first merge our datasets so that both the synthesized and unsynthesized versions of BasePay are included in the same dataset. Then we create a new variable, S, setting that equal to 0 if the observation comes from the confidential data, and 1 if it comes from the synthetic data. Then, using logistic regression, we fit a model and, for each observation, estimate $\hat{p}_i$—the probability of that observation being in the synthetic dataset. Finally, we compare the distributions of propensity scores across the confidential and synthetic datasets with the propensity-score mean-squared error: pMSE. High utility is noted by the model not being able to distinguish between the confidential and synthetic values, so a lower pMSE value indicates this high utility.

2.2.2 eCDF

The empirical cumulative distribution function, eCDF, is a discrete function which considers every observation in the sample to be an outcome of equal likelihood. In the context of synthesized data, a global

utility measure can be obtained by comparing the eCDFs of the synthetic and the confidential data. Ideally, they should be similar. Following Woo et al. (2009), for both the synthetic and the confidential datasets, we estimate the percentile of each record under the empirical cumulative distribution function. Two measures are calculated: U_m, the maximum absolute difference between the confidential and synthetic eCDFs; and U_a, the average squared differences between the confidential and synthetic eCDFs. As is intuitive, the smaller these values, the greater the utility of the synthetic dataset.

2.3 Analysis-Specific Utility

Analysis-specific utility focuses on whether similar statistical inferences can be obtained from the synthetic dataset as from the confidential dataset. These measures are generally dependent on what kind of analysis it might be expected that a data analyst would want to perform on the synthetic data. For our analysis-specific utility, we first check the inferences on both datasets for the mean and the regression coefficients. Then we evaluate the interval overlap using the results of these inferences.

2.3.1 Inference

Say one is interested in inferring a univariate parameter of interest from the synthetic dataset—we used the population mean. Using the point estimate and the variance estimate, both from the confidential dataset, one can calculate the point estimate and the variance estimate across m synthetic datasets. One can also calculate the degrees of freedom in order to compute the confidence interval of this parameter of interest. Comparisons can then be made between the parameter and confidence intervals of the synthetic and the confidential datasets. Something very similar can be done with the coefficients of, say, a linear regression.

2.3.2 Interval Overlap

We use definition one of interval overlap, proposed by Dreschler and Reiter (2009). This measure uses the lower and upper bounds of the confidence intervals for some parameter of interest, making use of the maximum lower bound across the synthetic and confidential datasets and the minimum upper bound across the synthetic and confidential datasets in order to give a measure to the overlap between these intervals. This definition is only measured from zero to one, where one indicates a high overlap and zero indicates that there is no overlap.

2.4 Risk Models

There are two types of disclosure which may occur with a synthetic dataset: identification disclosure, which occurs when an intruder identifies records of interest from the synthetic data; and attribute disclosure, which occurs when an intruder is able to infer the correct confidential value from the synthetic data. We first focus on identification disclosure risk evaluation, using the expected match risk and the false versus true match rates, and then the idea of record linkage; then we focus on attribute disclosure risk evaluation, using a classification-based risk measure.

2.4.1 Expected Match Risk

We use a matching-based approach to calculate both the expected match risk and the true/false match rate. For each record i, we assume that an intruder knows the true values for some of the covariates. Using this assumed knowledge, we calculate how many records in the synthetic data match with our record i. Then we can determine whether the true match is among these records, and how likely it is to find the correct match for each record. We use Bayesian probabilistic matching, from Reiter and Mitra (2009), though we use a more basic form.

The expected match risk is a measure of how likely it is to find, across the sample, the correct match for each record. It is a summation of fractions where the numerator is 1 if the true match is among those with the highest match probability, and 0 otherwise; and the denominator is the number of records with the highest match probability. The expected match risk can take a value from zero to n, where the higher the value in relation to your sample size, the higher the identification disclosure risk for the sample.

2.4.2 True/False Match Rate

The true match rate measures the size of the percentage of true unique matches. The higher the true match rate, which can be from zero to one, the higher the identification disclosure risk for the sample.

The false match rate measures the size of the percentage of unique matches which are false. The higher the false match rate, which can be from zero to one, the lower the identification disclosure risk for the sample.

2.4.3 Record Linkage

Record linkage methods were originally developed as a way to link records from multiple databases, but William E. Winkler (2004) established a method for applying record linkage to identification disclosure risk evaluation. These methods attempt to link the synthetic and confidential records, giving us the measures of true link percentage and false link percentage. The higher the true link percentage, the lower the false link percentage and the higher the identification disclosure risk, and vice versa.

Again, we make assumptions about which variables the intruder has information on. Using this information, we can generate pairs between the synthetic and confidential datasets. Then for each pair of records, we compare the values of the synthesized variables and create similarity scores. Using an expectation-maximization algorithm as proposed in Winkler (2000), we assign a weight value to each pair, and determine links. Finally, we link one-to-one the records from the confidential and synthetic datasets, and we can calculate the true and false link percentages.

2.4.4 Classification-Based Risk

Classification-based risk is an extension of the correct attribution probability (CAP) statistic, a measure that attempts to predict the value of a particular synthesized variable using some or all of the remaining variables. However, CAP uses a very simple model to do this, leading Choi et al. (2017) and Kaur et al. (2021), for example, to suggest the use of a more general classification method to evaluate attribute risk.

We use the synthetic dataset, up to all of the variables excluding the one for which we want to predict a value, and use a classifier to predict the value for that target variable. Then we do the same on the confidential dataset. We compare the mean-squared errors of these predictions, as well as calculate the proportion of observations in which the synthetic data has a less accurate prediction compared to the confidential data. For better attribute risk disclosure, we wish the mean-squared error of the synthetic data to be higher than the confidential; and we wish the proportion of observations in which the synthetic data has a less accurate prediction to be large.

2.4.5 Differential Privacy

Given the confidential nature of our dataset and specifically our outcome variable, protection is required for the disclose of summary statistics. Following the work from Dwork et al. (2006), differential privacy is a way to provide privacy protection for summary statistics. More specifically, the idea behind differential privacy is to add random noise dependent on an analyst's specified privacy budget to the output of summary statistics calculated from data. The release of the true value of summary statistics poses a breach of confidentiality, but by adding random noise, this risk is mitigated. We start by defining several key terms:

1. Database: A database is a confidential dataset.

2. Statistic: A statistic is any numeric attribute pertaining to a dataset that can be represented as a function, $f : \mathbb{N}^{|X|} \to \mathbb{R}_k$, that maps databases to k real numbers.

3. Hamming-distance: The Hamming-distance, represented by $\delta(x, y)$, is equal to $\#\{i : x_i \neq y_i\}$ where x and y are databases in $\mathbb{N}^{|X|}$. In the differential privacy setting, we are considering the scenario where two databases x and y differ by a singular record ($\delta(x, y) = 1$).

4. l_1-sensitivity: The l_1-sensitivity is a metric to gauge the "worst case scenario" denoted by Δf. Further, the l_1-sensitivity is the maximum change in the function f on x and y where $x, y \in \mathbb{N}^{|X|}$ and $\delta(x, y) = 1$. It is the case that sensitivity and added noise are positively related.

5. ϵ-differential privacy: The idea behind ϵ-differential privacy is to ensure that a mechanism acts similarly given the presence of noise on similar inputs. This is done by bounding (from above) the ratio of the difference in outputs from database x and database y of $\delta(x, y) = 1$ after the output has undergone some mechanism.

6. Privacy budget: The privacy budget is ϵ that was introduced in the previous definition. ϵ is specified by the data analyst and creates an upper bound for the differential privacy ratio. setting the maximum permissible difference between the log ratio of the probabilities of the outputs. It is the case that added noise and the privacy budget are negatively related.

With these understandings in place, we now introduce our two summary statistics of interest: mean BasePay and median BasePay. As mentioned above, we are considering the scenario where two databases x and y differ by one record (i.e. $\delta(x, y) = 1$). Therefore, x is our confidential Glassdoor sample and y is a database where one data entry differs from x. To calculate our summary statistics, we first find the mean and median of the confidential (unsynthesized) version of BasePay. When then specify our values of a, the lower bound for BasePay, b, the upper bound for BasePay, and n the number of observations in our database. The specifications of a and b depend on our own intuition. We then calculate the Δf for mean and median statistics using the following formulas: $\frac{b-a}{n}$ for mean and $b - a$ for median.

We now take a slight detour to introduce the Laplace distribution. The Laplace distribution follows a normal distribution with mean 0 and scale Δf: $X \sim$ Laplace $(0, \frac{\Delta f}{\epsilon}))$. The scale controls the spread, and thus, the noise; if more noise is needed, the scale is simply increased.

Shifting gears back to our summary statistics, we then specify our privacy budget value: $\epsilon = 0.1$. Following the properties of sequential composition, given we calculated two statistics from the same dataset, our privacy budget needs to be divided by two: $\epsilon_{new} = \frac{\epsilon}{2} = \frac{0.1}{2} = 0.05$. Using the Laplace distribution, we find the true BasePay average with added Laplace noise and the true BasePay median with added Laplace noise and compared those values to the true BasePay average and true BasePay median.

3 Results

Our results for all of our utility and risk evaluations can be found in Table 2. We used several different combinations of predictors in order to craft our synthetic data, but we focus on reporting the results of our final model, the one created by the two variables Age and Seniority. Additionally, we assumed several different combinations of variables that the intruder could know in order to provide an in-depth summary of the possible results. We focus on reporting the results of two different extremes: one where the intruder only knows Gender, and one where the intruder knows Gender, Age, and Dept.

Table 2: Utility-Risk Trade-off by Model Size

	Number of Predictors Included in Each Model			
Statistic	Two	Three	Four	Five
Adjusted R-Squared	0.5896	0.8186	0.8324	0.8412
pMSE	0.01384945	0.005782246	0.002505309	0.001189918
eCDF (Um)	0.852	0.616	0.636	0.595
eCDF (Ua)	0.03387857	0.01568713	0.01451311	0.01393157
Mean Evaluation Interval Overlap	0.9852073	0.9312128	0.9824421	0.9757288
Expected Match Risk (Syn Gender)	4.039313	5.47716	5.651184	5.631414
Expected Match Risk (Conf Gender)	7.799556	7.799556	7.799556	7.799556
False Match Rate (Syn Gender)	NaN	NaN	NaN	NaN
False Match Rate (Conf Gender)	NaN	NaN	NaN	NaN
True Match Rate (Syn Gender)	0	0	0	0
True Match Rate (Conf Gender)	0	0	0	0
Unique Matches (Syn Gender)	0	0	0	0
Unique Matches (Conf Gender)	0	0	0	0
Expected Match Risk (Syn Gender, Age, Dept)	372.3333	494.05	501.5667	510.2
Expected Match Risk (Conf Gender, Age, Dept)	657.6	657.6	657.6	657.6
False Match Rate (Syn Gender, Age, Dept)	0.3721591	0.2120419	0.1957672	0.2030848
False Match Rate (Conf Gender, Age, Dept)	0	0	0	0
True Match Rate (Syn Gender, Age, Dept)	0.221	0.301	0.304	0.31
True Match Rate (Conf Gender, Age, Dept)	0.411	0.411	0.411	0.411
Unique Matches (Syn Gender, Age, Dept)	352	382	378	389
Unique Matches (Conf Gender, Age, Dept)	411	411	411	411
True Linkage Percent	0.076	0.048	0.059	0.067
False Linkage Percent	0.924	0.952	0.941	0.933
kNN Syn MSE	437355049	280562345	279131052	274680293
kNN Conf MSE	260475523	260475523	260475523	260475523
Mean Where Rel Error of Conf > Rel Error of Syn	0.387	0.461	0.472	0.476

Looking at Figure 1, it's clear that some synthetic datasets are better fits than others; we largely evaluated upon the first dataset, excepting those cases in which we evaluated across all 20 that were generated.

3.1 Global Utility

3.1.1 pMSE

Our pMSE measure for the two-variable model is about 0.0138. This calculated propensity score is small and close to 0, suggesting our model can't distinguish between the confidential and synthetic datasets. The low score indicates a high level of utility of our synthetic data.

3.1.2 eCDF

The calculated empirical CDF utility measure of U_m is 0.852, which is close to 1. This suggests the maximum absolute difference between the confidential empirical CDF and the synthetic empirical CDF is large, which is not an ideal result. However, our U_a is about 0.0339, which is close to 0. This suggests the average squared differences between the confidential empirical CDF and the synthetic empirical CDF is small—so though our maximum difference is large, on average, the difference is much smaller. These indicate relatively high utility for our synthetic data.

3.2 Analysis-Specific Utility

3.2.1 Mean: Inference and Interval Overlap

In the confidential dataset, the mean is given as 94,472.65, with a 95% confidence interval from 92,900.34 to 96,044.96. Meanwhile, in our synthetic dataset, the mean is 94,425.28, with a 95% confidence interval

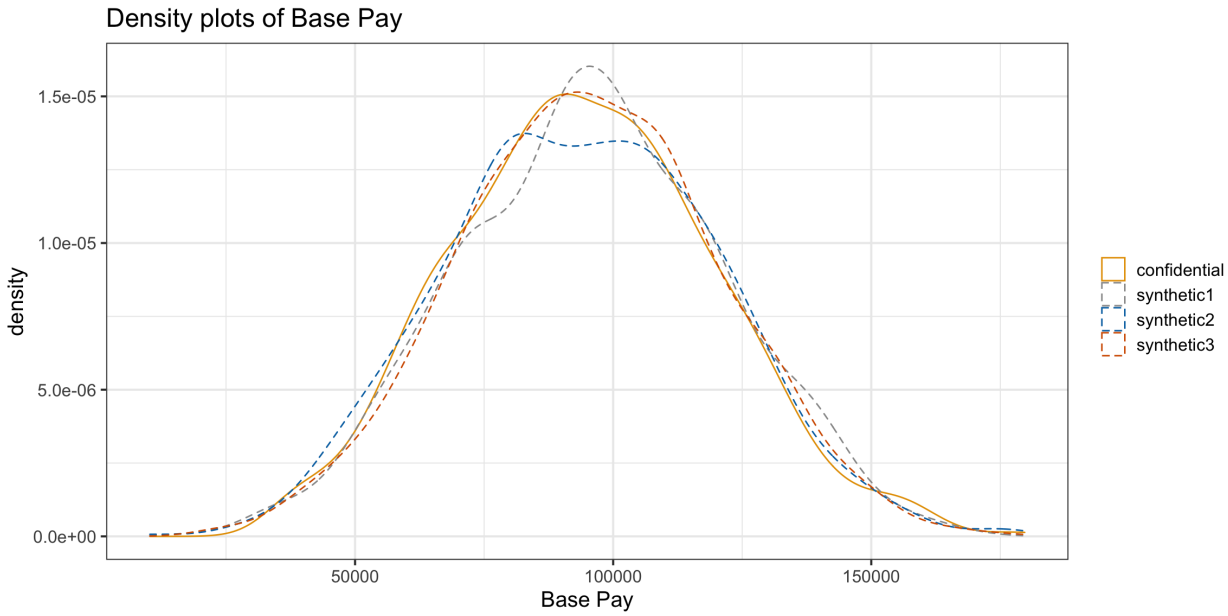

Figure 1: Plot of three synthetic BasePays and confidential BasePay

from 92,819.21 to 96,031.35. These results are all very close to one another, marking our interval overlap as 0.9852073. The closeness of this overlap to 1 indicates a high utility for mean inference.

3.2.2 Regression Coefficient: Inference and Interval Overlap

All of the regression coefficient results are very strong, with excellent overlap for the 95% confidence intervals of these as well. For example, Seniority5—the fifth category of the Seniority variable—has an overlap of .9812. Age has an overlap of .8982, and Seniority3 has an overlap of .9149.

3.3 Risk Evaluation

3.3.1 Identification Disclosure Risk

3.3.1.1 Expected Match Risk, True/False Match Rate:

Known Variables (Gender): When the known variable is only Gender (chosen because it seems like the most obvious variable for an intruder to know), the expected match risk on the confidential dataset is 7.800, compared to the synthetic dataset's 4.039. With 1,000 observations, there clearly isn't much risk overall even with the confidential data, but the synthetic data does reduce that by over three expected matches. Our false match rate and true match rates are both essentially 0, for both datasets, suggesting that no matches can be made, period, when Gender is the only known variable.

Known Variables (Gender, Age, Dept): When the known variables increase, our synthetic data's improvements compared to the confidential data's become more clear. We chose Gender again, as it is a straightforward variable to know; Age, as it is something we think an intruder may be able to find out about a person; and Dept, as the department categories are general and we think it's probable that an intruder would know this.

The expected match risk on the confidential dataset where these three variables are known is 657.6. This equates to correctly matching almost two-thirds of the data. Meanwhile, our expected match risk of the synthetic data is only 372.33—a reduction by about 43%, and equating to the correct matches being only just more than a third of the data. Similarly, while the false match rate of the confidential data is 0,

the false match rate of the synthetic data rises to 37.22%, a significant improvement. The true match rate also decreases from the confidential's 41.1% to the synthetic's 22.1%, a decrease by nearly half. The unique matches also fall from 411 to 352.

3.3.1.2 Record Linkage Our record linkage results are also very strong: only a 0.76% true match rate, equaling a 92.4% false match rate.

3.3.2 Attribute Disclosure Risk

While the MSE of our confidential data is 260,475,523, our synthetic MSE is 437,355,049. This suggests that the classification prediction of BasePay on our synthetic dataset is much less accurate compared to that of the confidential dataset, indicating a lower risk for the synthetic versus confidential data.

Furthermore, the mean where the relative error of the confidential data is greater than the relative error of the synthetic data is only 0.387, also indicating less accuracy—and therefore less risk—of the synthetic dataset over the confidential dataset.

3.3.3 Differential Privacy

The results from our differential privacy work are illustrated in Table 3. Taking the difference between our True BasePay mean and the True BasePay mean with Laplace noise, one can observe the true effect of the noise: $94,472.65 - 96,408.60 = -1,935.95$. In a similar vein, taking the difference between our True BasePay median and the True BasePay median with Laplace noise, one can observe the true effect of the noise: $93,327.50 - -2,284,298 = 2,377,625.50$. Given the continuous nature of our outcome variable, the extreme difference in True BasePay median and True BasePay median with Laplace noise is due to the way in which one calculates the Δf statistic for median (only $b - a$ without dividing by n). As such, no substantive conclusions should be made based on the median summary statistic. That said, through the use of added Laplace noise, we create a protected mean summary statistic that behaves similarly to the confidential mean summary statistic and eliminates the need and associated risk of releasing the true (confidential) mean summary statistic to the public.

Table 3: Differential Privacy Results

	Mean	Median
True BasePay	$ 94,472.65	$ 93,327.50
True BasePay with Laplace Noise	$ 96,408.60	$ –2,284,298

4 Discussion

As introduced briefly in section 2, we used best subsets to select our predictors for our final model. Looking at adjusted R-squared and considering the utility-risk tradeoff, we initially selected a model with four predictors (rather than our final model with two): JobTitle, Age, Education, and Seniority. After significant analysis, we found that our utility was incredibly strong yet our risk was quite poor.

Synthetic data with high utility and poor risk is very problematic when working with confidential data. Figure 2 compares the density plots of three synthetic BasePays and confidential BasePay for the two, three, four, and five predictor models. As one might assume (and can also see), plots of the synthetic BasePays and confidential BasePays are more aligned (i.e. the synthetic BasePays follow the shape and trend of the confidential BasePay better) in the models with more predictors compared to the models with less predictors. That said, even the worst-performing density plot (the two-predictor model) performs quite well.

Table 2 displays our risk measures by model size. One may notice that as the model size decreases (goes from five predictors down to two predictors), the risk improves greatly. That is, the expected match risk decreases, the true match rate decreases, the false match rate increases, and the number of unique matches decreases.

Given privacy and the protection of our confidential data is of the utmost priority, we ultimately decided

that sacrificing some utility for significantly improved risk and greater protection of the confidential data was worthwhile.

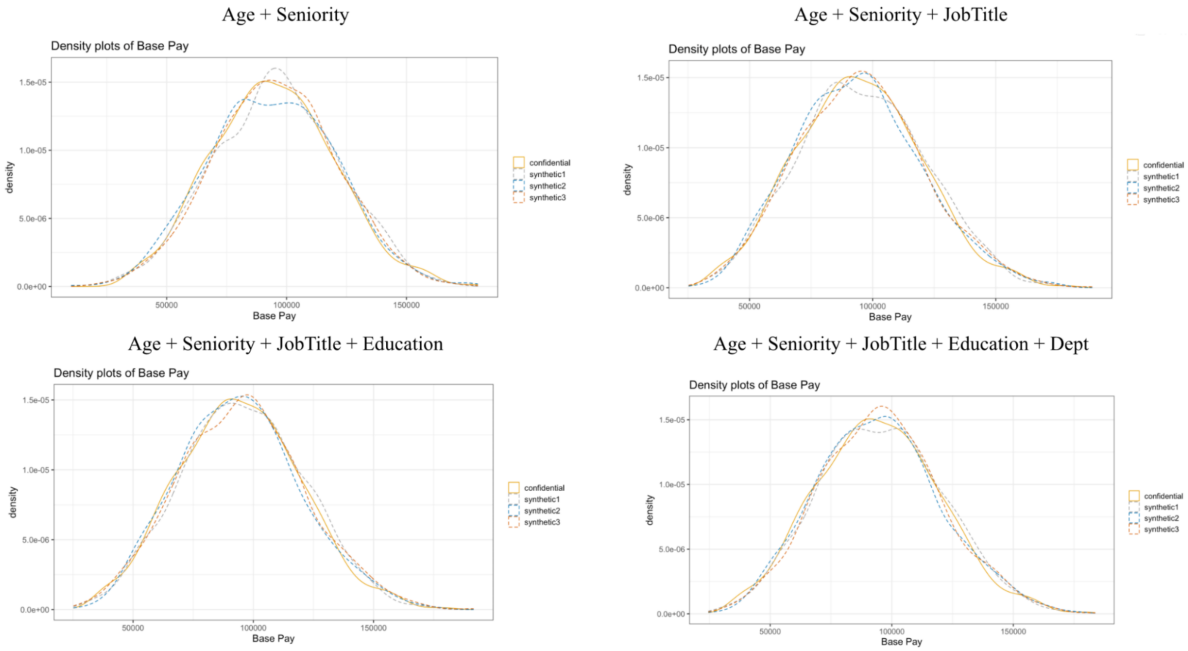

Figure 2: Comparison of Density Plots By Model Size

Based on our risk results given in section 3, we can say that our synthetic data provides sufficient privacy protection. Risk is significantly better in the synthetic data versus the confidential data under every measure: the expected matches decrease by over 40%, the true match rate decreases and the false match rate increases, the percentage of true linkages decreases and the percent of false linkages increases, the MSE of the synthetic BasePay is larger than the MSE of the confidential BasePay, and the synthetic data has a greater error than the confidential data over 60% of the time. While these risk results could be improved, it would definitely come at the expense of significant utility.

4.1 Further Research

Furthering the results shown here, the next step could be additionally synthesizing the demographic variables such as Age, Gender, and Education. This would further decrease the risk of both identification and attribute disclosure; however, it would likely also decrease the utility of the synthetic data. If this step is taken, perhaps a better model for BasePay might be used. Additionally, the order in which the variables should be synthesized would have to be decided, if a sequential synthesis were to be chosen.

5 References

Choi, Edward, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. (2017). "Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks." In Proceedings of the 2nd Machine Learning for Healthcare Conference, edited by Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, 68:286–305. Proceedings of Machine Learning Research. Boston, Massachusetts: PMLR.

Drechsler, J., and J. P. Reiter. (2009). "Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB." Journal of Official Statistics, 589–603.

Drechsler, J. (2011). Synthetic Datasets for Statistical Disclosure Control. Springer: New York.

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, 265–284.

Jauhari, N. (2019, February). Glassdoor- Analyze Gender Pay Gap, Version 1. Retrieved September 26, 2022 from https://www.kaggle.com/datasets/nilimajauhari/glassdoor-analyze-gender-pay-gap.

Kaur, D., M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, and N. Markuzon. (2021). "Application of Bayesian Networks to Generate Synthetic Health Data." Journal of the American Medical Informatics Association 28: 801–811.

Little, R. J. A. (1993). Statistical analysis of masked data. Journal of Official Statistics 9, 407–426.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. Journal of Official Statistics 19, 1–16.

Reiter, J. P., and R. Mitra. (2009). "Estimating Risks of Identification Disclosure in Partially Synthetic Data." The Journal of Privacy and Confidentiality 1: 99–110.

Rubin, D. B. (1993). Discussion statistical disclosure limitation. Journal of Official Statistics 9, 461–468.

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: 663–688.

Winkler, W. E. (2000). "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage." U.S. Bureau of the Census. Winkler, William E. 2004. "Re-Identification Methods for Masked Microdata." In Privacy for Statistical Databases, edited by J. Domingo-Ferrer and V. Torra, 216–230.

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: 111–124.