Privacy Protection of The Patient Characteristics Survey Using Bayesian Data Synthesis

Abstract

The necessity of synthetic data sets stems from the need for data analysis and processing coupled with the need for privacy and security. Without security, the public's private information is at risk. Additionally, the synthesizing and randomizing of data, but without statistically accurate data, can lead to the inability to preform proper data analysis. Through a number of measures, we explore and evaluate the utility and risk of such a method of synthesizing data through the Dirichlet Process Mixture of Products of Multinomial distributions model on a patient characteristics survey.

1 Introduction

Many companies, policy makers, and healthcare professionals make their important decisions by analyzing data. The ability to make good decisions, however, becomes limited when data is also limited. This is inevitable as this data often risks compromising the privacy of the individuals they represent. To circumvent the lack of data, many researchers have sought synthetic data sets which make the variables of a data set untraceable from the individual it was taken from and thus protecting the individuals while providing data for important decision making.[1] However, we also have to consider that if the variables are made too different from what they originally were, we risk losing statistical utility. That is, the hidden statistical inferences and connections between the variables may be lost if we simply randomize everything and thus releasing the data set would be meaningless as you can't make important decisions based on random, statistically inaccurate data. To strike the balance between randomization for privacy protection and preservation for statistical analysis we trained and extensively tested a DPMPM synthesis model [3] on public health data.

1.1 The PCS data sample

Health data and information are an important tool for research. As much of this research is used to promote high quality health and protect the public's well-being, the U.S. Department of Health and Human Services issued the *Standards for Privacy of Individually Identifiable Health Information*[2], or the Privacy Rule, in 2000 as a means for this information to be collected while also properly protecting individual's health information. The Privacy Rule addresses the standards for individual's rights to control and understand how their health information is being used. We aim to ensure that individuals in health data have their privacy protected while also transferring information that continue to nurture good health policy and advancements.

The health data we will be using is from the Patient Characteristics Survey (PCS) of 2019. The PCS is a data source that collects specific client-level information from all public health service programs in New York State's public health system. There are 140,875 patients included in this sample and 75 categorical variables ranging from demographics to health information. For this analysis, a randomized sub sample of 5000 patients was generated and we only selected 11 variables which include gender, employment status, and drug substance disorder. All variables we selected, including what variables we decided are sensitive, and thus unsuitable to release their true values in respect for a patients privacy, can be viewed from Table 1.

The rest of the paper will be organized as follows. Section 2 will go into depth about the DPMPM model, how it works, and how it is implemented. Section 3 will go through the various utility and risk evaluations we put our synthetic datasets through. And Section 4 will conclude with the interpretation and implications of our results as well as possible future work.

Table 1: Selected Variables from the Patient Characteristics Survey

Non-Sensitive	Sensitive
Age (binary[1])	Region served (categorical[5])
Gender (binary)	Alcohol-related disorder (binary)
Race (categorical[2])	Drug substance disorder (binary)
Education (categorical[3])	Cannabis use (binary)
Health insurance (binary)	SSDI Cash Assistance (binary)
Employment status (categorical[4])	

[1] Age is split into 1-Child and 2-Adult. Gender is split into 1-Male and 2-Female. All other binary variables are split into 1-no and 2-yes.

[2] Race is categorized into 1-Black Only, 2-Multi-Racial, 3-Other, 4-White Only.

[3] Education is categorized into 1-No formal Education, 2-Pre-K to Fifth Grade, 3-Middle School to High School, 4-Some College, 5-College or Graduate Degree.

[4] Employment status is categorized into 1-Employed, 2-Non-Paid/Volunteer, 3-Unemployed and not looking for work, 4-Unemployed, looking for work.

[5] Region is categorized into 1-Central NY, 2-Hudson River, 3-Long Island, 4-New York City, 5-Western.

2 The DPMPM synthesizer

The DPMPM is a joint approach to synthesizing multivariate categorical data. It was implemented by Hu et al. in 2014[3] to generate fully synthetic data, comprised of categorical variables, from the American Community Survey. Now we use it to synthesize our sensitive variable for the PCS.

Let's consider our confidential sample Y which consists of 5000 records. Each record $i = 1, ..., 5000$ has 5 unordered categorical variables. These variables are what we are interested in synthesizing. To use the DPMPM, we must assume that every record $Y_i = (Y_{i1}, ...Y_{i5})$ belongs to one of K underlying unobserved latent classes. For indication of the class assignment for record i, denoted z_i, let $\pi_K = Pr(z_i = K)$. We assume that $\pi = (\pi_1, ..., \pi_K)$ is the same for all records. For any $c \in \{1, ..., d_j\}$, where d_j is the number of categories in variable j and $j = 1, ..., 5$, let $\theta^{(j)}_{z_i d_j} = Pr(Y_{ij} = c \mid z_i = K)$. Let θ be the collection of all $\theta^{(j)}_{z_i d_j}$. Given the latent class assignment z_i of record i and θ, each variable Y_{ij} independently follows a multinomial distribution. The distributions can be expressed as

$$Y_{ij} \mid z_i, \theta \overset{ind}{\sim} \text{Multinomial}(\theta^{(j)}_{z_i 1}, ..., \theta^{(j)}_{z_i d_j}; 1) \ \forall \ i, j \tag{1}$$

$$z_i \mid \pi \sim \text{Multinomial}(\pi_1, ..., \pi_K) \ \forall \ i. \tag{2}$$

The marginal probability of $Pr(Y_{i1} = y_{i1}, ..., Y_{ip} = y_{ip} \mid \pi, \theta)$ can be expressed as

$$Pr(Y_{i1} = y_{i1}, ..., Y_{ip} = y_{ip} \mid \pi, \theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{r} \theta^{(j)}_{k y_{ij}} \tag{3}$$

where it is averaging over the latent classes.

Priors need to be provided for the parameters, π and θ, in the DPMPM. First, we select a large enough K, the upper bound of possible latent classes to be used, to allow the DPMPM to fully explore the parameter space and determine the effective number of occupied latent classes. We used $K = 70$ latent classes. To empower the DPMPM to pick an effective number, the truncated stick-breaking representation[4] of the Dirichlet Process prior is used. It is expressed as

$$\pi_k = V_k \prod_{I<k} (1 - V_I) \text{ for } k = 1, ..., 70, \tag{4}$$

$$V_k \overset{iid}{\sim} \text{Beta}(1, \alpha) \text{ for } k = 1, ..., 70 - 1, \ V_K = 1, \tag{5}$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha), \tag{6}$$

$$\theta_k^{(j)} = (\theta_{k1}^{(j)}, ..., \theta_{kd_j}^{(j)}) \sim \text{Dirichlet}(a_1^{(j)}, ..., a_{d_j}^{(j)}) \text{ for } j = 1, ..., 5, \ k = 1, ..., 70. \tag{7}$$

To allow the data to dominate the prior distribution, we set $(a_\alpha = 0.25, b_\alpha = 0.25)$. The posterior distribution of all parameters is estimated through using a blocks Gibbs sampler[6]. For data synthesis, the confidential data set is used for model estimation through MCMC. For our model, we did 5000 iterations. At chosen MCMC iteration I, we first sample a value of the latent class indicator z_i (2). Given the sampled z_i, we sample synthetic values of sensitive variables using independent draws (1). This process is repeated for every record that has sensitive values to be synthesized until we obtain one synthetic data set. To create $m > 1$ synthetic data sets, we repeat the process of using independent draws of the parameters and sampling synthetic values m times. Let $S = (S^{(1)}, ..., S^{(5)})$ denote the set of five synthetic categorical data sets generated for our purposes.

To note, the NPBayesImputeCat package in R was used to create the DPMPM for this analysis.[5]

While there are many methods of data synthesis, since all of our variables are categorical we chose to explore the Dirichlet process mixture of products of multinomial distributions model (DPMPM). To show consistency, our measures are either 5 or the average of 5 DPMPM, synthetically generated datasets.

3 Utility and Disclosure Risk Evaluation Results

3.1 Utility Evaluation

We evaluate the preservation of statistical utility from our confidential dataset to the synthetic dataset. To do so we set up an array of statistical measures that quantify the relationship between the variables and we compare the quantities to confirm that they stay relatively the same once they are synthesized.

3.1.1 Propensity Score Mean-Squared-Error

Propensity scores are a measure that represent the probability for individuals in a dataset to be assigned a specific treatment group given their information on other variables. We can use this as a measure of utility by using it to investigate whether the synthetic observations significantly differ from the original observations. In other words, we will measure the pMSE of the observation being part of the synthetically generated dataset. For this measure we follow Woo et al. (2009)[7] and Snoke et al. (2018)[8].

To calculate this measure we explored the use of two classification methods, logistic regression and multi-layered perceptrons (mlp).[9] Both models have all of our variables as the input values, on a dataset that merges the confidential and synthetic data and the model is trained to predict an extra, added variable indicating whether the data came from the confidential or synthetic dataset.

The average pMSE over our 5 synthetic datasets when trained with logistic regression is 0.00561482. In comparison, the average pMSE over our 5 synthetic datasets when trained with a multi-layered perceptron AI is 0.006716219. The average pMSEs are small and close to 0 meaning the classification models cannot distinguish between the confidential and synthetic datasets, indicating a high level of utility on our synthetic datasets.

3.1.2 Distributions of Differences in Relative Frequencies

Another measure of global utility is adopted from Drechsler and Hu (2021)[10]. After computing the relative frequencies of each cell in various cross tabulations of only our sensitive variables, such as marginal distributions, two-way interactions, and three-way interactions, we assess how these frequencies differ between the confidential and synthetic data. We calculate the relative difference in frequencies as

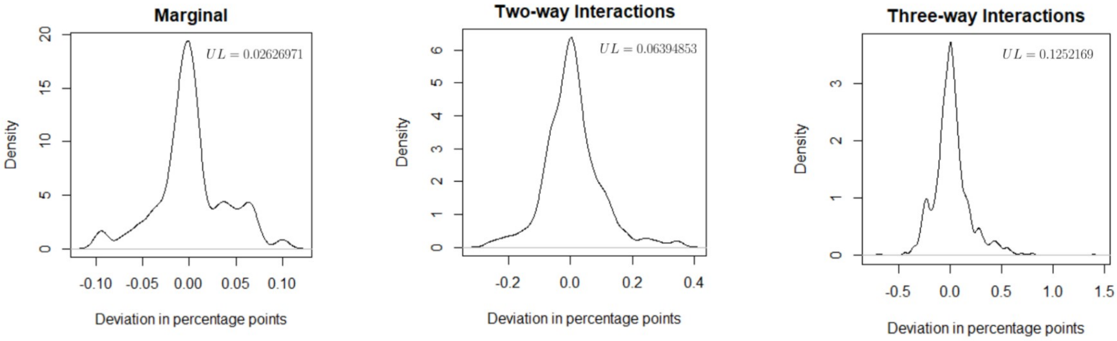
$$\text{Difference D} = \frac{\text{synthesized frequency-confidential frequency}}{\text{confidential frequency}}. \tag{8}$$

We also calculate a measure for the loss of utility as the average of the absolute value of all differences, n, which can be expressed as

$$UL = \frac{1}{n}\sum_{f=1}^{n}|D|. \tag{9}$$

We plot the distributions in Figure 1, with each distribution's respective utility loss displayed on the upper-right corner.

Figure 1: Difference in Relative Frequencies Distributions



Higher density remains around zero as we increase interactions indicating high global utility, although the density does decrease and deviations get wider as we increase interactions. This is evident from the utility loss measure which increases in each panel, but for the most part, remains closer to zero.

3.1.3 Proportions of Analysis Interest

Now we want to evaluate how different the statistical inference on the confidential data set is from the synthetic data set. To analyze analysis-specific utility, we compare the inference on population proportions for both data sets. We first calculate the point estimate and variance of the populations proportions of each of our sensitive variables from the confidential data set Y. Calculations follow the Central Limit Theorem for categorical variables. Then we calculate the average point estimates and variance from our synthetic data set S. Additionally, we obtain 95% confidence intervals for all point estimates.

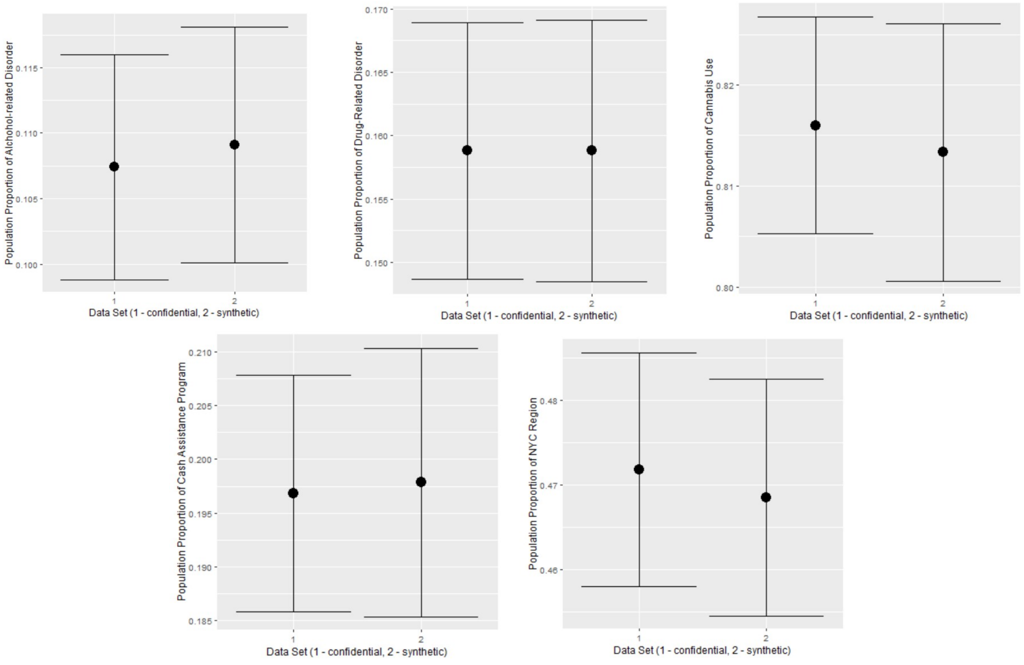
Moreover, we calculate the interval overlap for confidence intervals. The measure we use is given by Snoke et al. (2018)[8]. The interval overlap measure is calculated as

$$IO = \frac{1}{2}(\frac{min(U_c, U_s) - max(L_c, L_s)}{U_c - L_c} + \frac{min(U_c, U_s) - max(L_c, L_s)}{U_s - L_s}) \tag{10}$$

where (L_c, U_c) and (L_s, U_s) denotes the 95% confidence intervals for the confidential and synthetic data sets, respectively. The closer IO is to 1, the higher the utility. On the other hand, the further IO is to 1, the lesser the overlap, and the lower the utility. This measure allows IO to be negative. The bigger the absolute value of a negative IO, the further away the two intervals are.

We can evaluate differences between the confidential and synthetic data sets using Figure 2. Exact estimates are in Table 2.

Figure 2: Point Estimate and Confidence Intervals for Population Proportions



As we can see, point estimates of population proportions for sensitive variables are close to each other and confidence intervals have high overlap. This indicates that synthetic data sets have high analysis-specific utility.

Table 2: Point Estimate and Confidence Intervals for Population Proportions

		Point Estimate	2.5%	97.5%	Interval Overlap
Confidential	alc	0.1074	0.09881582	0.11598418	
	drug	0.1588	0.1486669	0.1689331	
	cann	0.816	0.8052571	0.8267429	
	cash	0.1968	0.1857772	0.2078228	
	NYC	0.4718	0.4579597	0.4856403	
Synthetic	alc	0.10908	0.1003487	0.1178113	0.9050747
	drug	0.1588	0.1486263	0.1689737	0.9916403
	cann	0.81332	0.800584	0.826056	0.8922849
	cash	0.19784	0.1853604	0.2103196	0.9416327
	NYC	0.46852	0.4545112	0.4825288	0.8822547

3.1.4 Regression Coefficients

Another measure of analysis-specific utility is regressing variables in the synthetic and confidential data set and comparing the regression coefficients and confidence intervals of these coefficients. We included interval overlap measures again in this analysis.

Using logistic regression, we regress our sensitive variables, denoted Z, on their statistically significant, at a 5% confidence level, predictors, denoted X_n. Results are in Table 3.

We can see that the synthetic data sets do a good job of preserving the relationships between variables, but not so well at capturing the confidence intervals of the confidential regression coefficients, as evidenced by the variation in interval overlaps. We can gather that there are some limitations to making inference by regressing on the synthetic data sets. Given we only regressed the sensitive variables by their statistically significant predictors, we can not evaluate the other possible regressions that do not regress the sensitive variables in the data.

3.2 Risk Evaluation

We evaluate the risk our synthetic dataset creates by simulating how a malicious third party would try to retrieve sensitive information from our data. For this evaluation we explore identification and attribution risk. Identification risk refers to the risk that the intruder can correctly identify the records of a person of interest from the synthetic data. Attribution risk refers to the risk that an intruder can correctly infer the true confidential values using information from our synthetic data.

3.2.1 Matching-based Approach

In this approach we use Bayesian probabilistic matching, a basic version of Reiter and Mitra (2009)[11]. We use three widely-used summaries of identification disclosure risks: the expected match risk, the true match rate, and the false match rate.

The expected match risk measures on average how likely it is to find the correct patient record. It is defined as: $\sum_{i=1}^{n} \frac{T_i}{c_i}$, where $T_i = 1$ if the true match is among

Table 3: Regressing Sensitive Variables on Statistically Significant Predictors

	Z	X_n	Point Estimate	2.5%	97.5%	Interval Overlap
Confidential	alc	age2	3.5271	2.3852276	5.327423	
		gender2	-0.6965	-0.9168425	-0.478833	
		drug2	2.6332	2.4239998	2.845800	
	drug	age2	2.21417	1.695321	2.8262401	
		gender2	-0.56453	-0.754525	-0.3756542	
		cann2	-1.71194	-1.908141	-1.5167045	
		alc2	2.6111	2.387762	2.8381538	
	cann	age2	-0.59139	-0.836191	-0.3564178	
		cash2	0.49687	0.2937945	0.7053863	
		drug2	-1.79426	-1.9658585	-1.6233478	
	cash	age2	2.64689	2.2894154	3.0432230	
		gender2	-0.3216	-0.4696509	-0.1738853	
		cann2	0.41576	0.2179116	0.6188615	
		emp2	1.55897	0.967344	2.1358404	
		emp3	1.07209	0.8591314	1.2927836	
		emp4	0.74518	0.4714102	1.0207696	
		insur2	2.10919	1.5850665	2.7250304	
	region	insur2	0.5432	0.2293981	0.8403632	
Synthetic	alc	age2	2.941288	1.793712	4.088864	0.6606559
		gender2	-0.26733124	-0.49243656	-0.04222591	0.03063684
		drug2	2.574244	2.346576	2.801911	0.8629557
	drug	age2	2.088164	1.547499	2.628828	0.8443685
		gender2	-0.7572185	-0.9448845	-0.5695525	0.4905221
		cann2	-1.324577	-1.580211	-1.068944	0.1432268
		alc2	2.445137	2.209735	2.680540	0.6359593
	cann	age2	-0.6247427	-0.867642	-0.3818433	0.9411322
		cash2	0.3300534	0.1057027	0.5544041	0.6069917
		drug2	-1.544498	-1.769118	-1.319878	0.3750375
	cash	age2	2.296363	1.842002	2.750724	0.5598084
		gender2	-0.11687709	-0.27258582	0.03883165	0.3253258
		cann2	0.22598385	0.01631325	0.43565445	0.5311587
		emp2	0.9171959	0.1858136	1.6485783	0.5243589
		emp3	0.8162837	0.5872915	1.0452759	0.4178456
		emp4	0.6205783	0.2948454	0.9463113	0.7967182
		insur2	1.3339728	0.8640567	1.8038889	0.2123953
	region	insur2	0.5068919	0.1458955	0.867882	0.9231103

the c_i units and $T_i = 0$ otherwise, and c_i is the number of records with the highest match probability for the target record i. The higher the expected match risk, the higher the identification disclosure risk for the sample, and vice versa.

The true match rate represents the percentage of true unique matches that exist. It is defined as: $\sum_{i=1}^{n} \frac{K_i}{N}$, where $F_i = 1$ if there is a unique match but it is not the true match (i.e., $c_i(1 - T_i) = 1$) and $F_i = 0$ otherwise, and s is the number of uniquely matched records (i.e., $\sum_{i=1}^{n}(c_i = 1)$), and N is the total number of target records. The lower the true match rate, the higher the identification disclosure risk for the sample, and vice versa.

The false match rate represents the percentage of unique matches that are actually false matches. It is defined as: $\sum_{i=1}^{n} \frac{F_i}{s}$, where $K_i = 1$ if the true match is the unique match (i.e., $c_i T_i = 1$) and $K_i = 0$ otherwise, and N is the total number of target records out of n total records. The lower the false match rate, the higher the identification disclosure risk for the sample, and vice versa.

Table 4: Confidential Identification Disclosure Risk Summaries

	Confidential	Synthetic
Expected Match Rate	453	27.07142
True Match Rate	0.0246	0.00044
False Match Rate	0	0.9833207

Assuming the intruder knows the age, gender, and race of their targeted, confidential record, we want to see which records in the synthetic data are matched with the confidential record. As we can see from Table 4 our synthetic true match rate is low and false match rate is high, especially compared to the confidential records. This means it is not very likely to find the correct match for each record, indicating a low identification risk.

3.2.2 Record Linkage Approach

For this method we link records in the synthetic dataset to the records in the confidential dataset and among these linkages, we can evaluate identification risks in terms of true links (i.e. correct links) and false links (i.e. incorrect links). Following the general procedure from Fellegi and Sunter (1969)[12], by simulating an intruder that's been given an individual's age, race and gender we generate pairs between the synthetic and confidential dataset. For each pair we then compare the values of the variables the intruder was not given and calculate a similarity score. We then select one-to-one linkages between the synthetic and confidential dataset and use the pairs and their similarity scores to calculate the percentages of true links and false links.

Table 5: Average Greedy Synthetic Data

	False	True
False	3340593	4855.6
True	4947.4	52.6

Table 5 and Table 6 depict the results of the record linkage approach from the synthetic and confidential data (Note: greedy represents the name of the algorithm used to calculate the weight or similarity score for each pair). The average true

Table 6: Greedy Confidential Data

	False	True
False	3340593	3780
True	3780	1220

linkage of our synthetic datasets is 52.6/5000 = 1.052%. The average false linkage of our synthetic datasets is 4947.4/5000 = 98.948%. As we have a low percentage of true links and a high percentage of false links our model thus shows low identification disclosure risk.

3.2.3 Correct Attribution Probability Measure

To measure attribution risk, we used the correct attribution probability (CAP) measure. The CAP measure was first proposed by Elliot (2014)[13] and Taub et al. (2018)[14]. It measures the probability that an intruder can correctly predict the value of the target for an individual, by using the empirical distribution of this variable among synthetic observations with the same key variables. We want to focus on record-level individual CAP meaning we want to focus on the probability that an intruder can correctly predict the value of the target (sensitive variable) for each of our records. We will only consider one synthetic data set here.

Consider a specific sensitive variable, I, that we synthesized. All possible values for this variable are targets and are denoted as $T_1, ..., T_G$. The intruder is attempting to predict the value of y_{iI} using some or all of Y^{-I}, the set of variables other than I in Y. The possible combinations of these variables are called keys, and they are denoted as $K_1, ..., K_H$. Each record, y_i, in the confidential data set is thus associated with single key $K(y_i)$ and a single target $T(y_i)$. We consider the same keys and targets in the synthetic data set S as in the confidential data set. The individual CAP of record y_0 in confidential data set Y with synthetic data set S is calculated as

$$CAP_{y_0}(S) = \frac{\sum_{i=1}^{5000} I[T(z_i) = T(y_0), K(z_i) = K(y_0)]}{\sum_{i=1}^{5000} I[K(z_i) = K(y_0)]} \tag{11}$$

if $\sum_{i=1}^{5000} I[K(z_i) = K(y_0)] \neq 0$ and 0 otherwise.

We create scatterplots (Figure 3 and Figure 4) comparing the individual CAP for each record before and after the synthesis process, but only for alcohol-related disorder and cannabis use as we think they are worth noting.

Note that patients who have an alcohol-related disorder are at a lower risk of having an intruder predict this true value. As these are the patients we want to protect this information for, we believe the synthetic data set does a reasonable job of lowering attribute risk for them. On the other hand, patients who use cannabis are at a higher risk of having an intruder predict this behavior. This could be explained by the fact that the majority of patients, about 80%, in the sample use cannabis. Therefore, the synthetic data set does not lower attribute risk for this particular group of patients.

3.2.4 File-Level Mean-Squared-Error

Our first classification-based measure measures the mean-square error of the predictions from our model trained on the synthetic and confidential datasets This section

Figure 3: Record-level Individual CAP for Alcohol-related Disorder

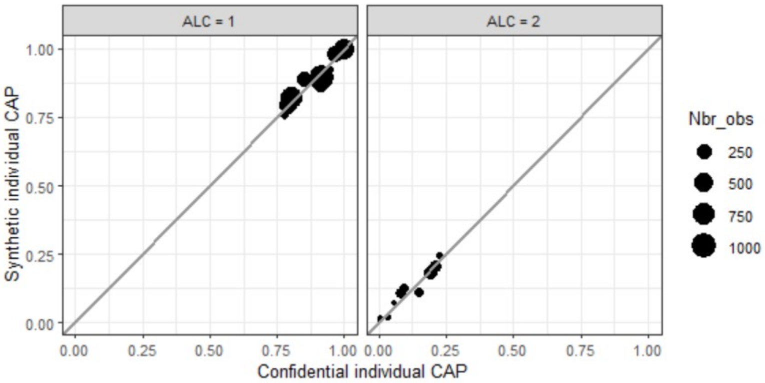
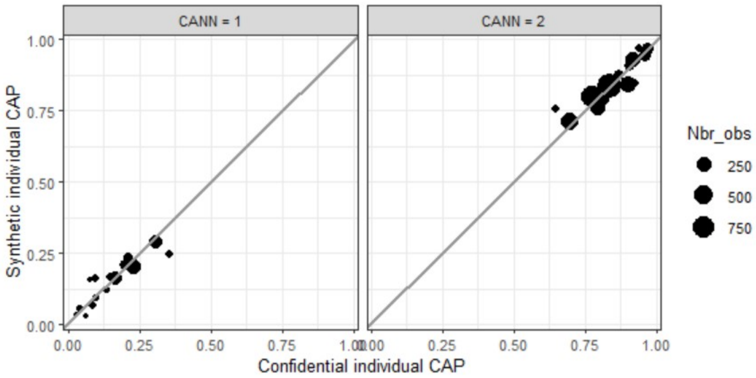


Figure 4: Record-level Individual CAP for Cannabis Use



takes inspiration from more general classification approaches to evaluating attribute disclosure risk in Choi et al. (2017)[15] and Kaur et al. (2021)[16]. We assume the intruder has information about the patient's age, gender, and race and will be looking for the real values of our sensitive variables from Table 1. Ideally, we would want a higher MSE for our synthetic datasets compared to our confidential datasets as that tells us it's more difficult to predict our sensitive variables using our synthetic dataset. MSE is calculated with the following formula where Y represents our confidential data and $\hat{Y}_i$ represents our model's prediction:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Table 7: Mean-Squared-Error with kNN and mlp

	Confidential MSE	Synthetic MSE
k-Nearest-Neighbors	6.971778	7.137289
Multi-layered Perceptron	1.659476	1.664963

For this measure we calculated MSE using two models to show consistency, a k-Nearest-Neighbors model and a Multi-layered Perceptron model. The results of calculating these measures can be found on Table 7, the synthetic MSE is consistently larger than the confidential MSE in all datasets and models indicating a low attribution risk.

3.2.5 Record-Level Prediction and Relative Error

Our second classification-based measure measures individual prediction errors. We first calculate the record-level relative absolute prediction errors in the confidential and the synthetic datasets. Using this information we return the proportion of observations that will have a less accurate prediction with the synthetic data compared to the confidential data. For each observation and variable relative error is calculated as, $\frac{|true\_value - predicted\_value|}{true\_value}$.

More simply we found the probability that confidential relative error was greater than the synthetic relative error and we hope to find this proportion to be low as a lower confidential relative error in comparison to synthetic would mean low attribution risk.

Using the same models from our MSE measure, we calculated the relative error of both the confidential and synthetic datasets. When compared we found that kNN tells us that 16% of patients are at a bigger risk of having their data being correctly predicted, this proportion rose to 57% using an mlp model. While this high proportion does indicate high attribution risk we also need to consider that our data is entirely categorical and contains many binary variables. This could mean our data is easy to guess or predict compared to something such as a continuous variable that has a wide range of possible values. The format of our data could mean high attribution risk in nearly every scenario. To test this we would need to test multiple synthetic datasets, especially a fully-synthetic dataset in order to drastically lower attribution risk, in order to have a relative idea of what a good proportion with this data format would be.

4 Conclusion

In terms of data synthesis, while our data and model passes most tests we found a fairly high attribution risk. As our utility is promisingly preserved we may benefit from slightly sacrificing it in favor of synthesizing more variables in order to reduce risk. For instance, if we were to synthesize the variable an intruder would most likely know it would become even more difficult for the intruder to find their intended individuals. However, we also have to consider that this high attribution risk can also be the result of the data being mostly binary. If an intruder wants to know the true value of a binary variable they already have a 50% of correctly guessing it. While we can work to lower attribution risk for this model we also need to consider that this model may already be sufficient given this context.

In terms of testing, we found that for the tests based in classification that artificial intelligence provided a rigorous critique. While other types of measures and model show relative promise for our model, our AI models consistently outperformed other models and has even showed an significantly high attribution vulnerability that would have otherwise gone undetected. While future work for our data synthesis model is needed we have learned that AI testing is essential for any iteration of our model going forward.

References

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

[2] (OCR), O. for C. R. (2021, July 27). Summary of the HIPAA privacy rule. HHS.gov. Retrieved December 8, 2021, from https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.

[3] Hu, J., Reiter, J. P., and Wang, Q. (2014). "Disclosure risk evaluation for fully synthetic categorical data." In Domingo-Ferrer, J. (ed.), Privacy in Statistical Databases , 185–199. Springer

[4] Sethuraman, J. (1994). "A Constructive Definition of Dirichlet Priors." Statistica Sinica 4: 639–50.

[5] Hu, J., Akande, O., Wang, Q. (2021). Multiple Imputation and Synthetic Data Generation with the R package NPBayesImputeCat. arXiv [stat.CO]. Opgehaal van http://arxiv.org/abs/2007.06101

[6] Ishwaran, H., and L. F. James. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." Journal of the American Statistical Association 96: 161–73.

[7] 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 (1): 111–24

[8] 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 (3): 663–88.

[9] Bergmeir C, Benítez JM (2012). "Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS." Journal of Statistical Software, 46(7), 1–26. https://www.jstatsoft.org/v46/i07/.

[10] Drechsler, J., and Hu, J. (2021). "Synthesizing geocodes to facilitate access to detailed geographical information in large scale administrative data". Journal of Statistics and Survey Methodology, 523–548.

[11] Reiter, J. P. 2005. "Estimating Risks of Identification Disclosure in Microdata." Journal of the American Statistical Association 100: 1103–12

[12] Fellegi, I. P., and A. B. Sunter. 1969. "A Theory for Record Linkage." Journal of the American Statistical Association 64 (328): 1183–1210

[13] Elliot, M. (2014). "Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team." CMIST.

[14] 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.

[15] 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.

[16] 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 (4): 801–11.