Using Synthetic Microdata to Maintain Privacy in Health Data: A NHIS Synthesis

Sarah Boese

May 6, 2020

Abstract

In this analysis, we look at a small subset of the National Health Interview Survey (NHIS) for 2018 pertaining to chronic conditions, worker status and missed work days. We establish that information contained in this data can make survey respondents vulnerable to identification by intruders. We can use Bayesian models and a method similar to posterior prediction to create new or synthetic micro-level data. This way, we can provide a greater level of privacy for participants in the NHIS survey while also maintaining high utility for the end user. In our Bayesian models, we use demographic data as unsynthesized predictors for this data. The effectiveness of this process demonstrates the correlation between this demographic data and the chronic conditions, worker status and missed work variables. It also demonstrates the power of Bayesian Logistic Regression. We find that our sythesizers work to yield high utility while demonstrating as sufficiently small amount of risk for participants in the NHIS survey.

1 Introduction

1.1 Research Questions

The National Health Interview Survey (NHIS) is a yearly survey conducted by the US Census Bureau concerning a broad range of health topics. Our research is focused on the Sample Adult 2018 file, just one set of microdata released each year. This particular data set includes 742 variables, many of which are sensitive. For our purposes, however, we limit our scope to only six variables: SEX (gender), RACERPI2 (race), AGE_P (age), FLA1AR (functional limitation), DOINGLWA (employment status) and WKDAYR (lost work days last year)[2]. We aim to answer the following questions: Are these demographic variables good predictors for FLA1AR and DOINGLWA? How can we best preserve the relationship between categorical variables describing functional limitation and work status to the continuous variable recording number of lost work days in a year? Moreover, does the entire synthesis process preserve to a satisfactory degree the "usefulness" of the data in further analysis? After we develop these models, the question becomes if such models still satisfactorily minimize risk of disclosure for participants.

1.2 Background/Significance of Research

Under Title 13, it is illegal for the US Census Bureau to "disclose or publish any census or survey information that identifies an individual or business". To that end, most data released by the US Census Bureau undergoes some conversion to ensure privacy. One way to protect participant's privacy is to release summary level data. However, releasing such data still requires the addition of noise or perturbation to the data in order to ensure privacy. This is the focus of research in the field of differential privacy **cite here**.

However, many analysts looking to use the data released by the US Census Bureau and other institutions wish to use micro-level instead of just summary level data. *Synthetic Microdata* is another research area looking to maintain privacy guarantees while also increasing utility for end users. *Microdata*, also called record-level or respondent-level data, is a collection of individual or business respondent data for a set list

of variables/attributes. We generate synthetic, or new, micro-data using a process similar to that used in Bayesian posterior predictive checks: first, we fit our data to pre-determined models using Monte Carlo Approximation. Then, we use a draw from our posterior chain to calculate new values for our columns of interest. Here, we may run into the possibility that our model works too well and we do not introduce enough noise to ensure the privacy of survey participants.

To measure this possibility, we must formalize what we mean when we say disclosure risk. *Disclosure Risk* is the risk posed by an intruder or attacker who uses a publicly available database to derive confidential information about individuals in the database. We will consider two types of disclosure risks: identification and attribute disclosure. *Identification Disclosure* is an intruder using mircrodata to find out about the identity of an individual that the intruder is specifically looking for. We say this is like your neighbor finding out information about you by identifying you in a dataset. *Attribute Disclosure*, by contrast, consists of an intruder correctly inferring the true value of one or more unknown variable(s) in an individual's response. We will use measures to measure these risks before and after synthesis to see if we have introduced privacy protection to the synthetic micro-level data[3].

Competing with the need to maintain privacy in released data, it is important that we release usable or high quality data. That is, it is pertinent that we maintain those relationships between different response variables that would be important analysts when using the data. Formally, we measure the *Utility* of the synthesized data, or quality of what can be learned about the original data from the synthesized data, by using *Utility Measures*. These can either be specific to our analysis, for example comparing distributions and confidence intervals of our original and synthetic data, or universal measures that can be used on any pair of original and synthetic data sets[5].

2 Methods Used to Obtain Data

We contain our focus to just six of the 742 variables in the NHIS Sample Adult data set. We also contain ourselves to a 1,000 observation subset of the total data which has 25417 records. Three of our variables of interest, namely those recording gender, race and age, we will use as predictor variables. The other three, those recording functional limitation status, work status and lost work days will be actually synthesized. We do this because, to our understanding, AGE and RACEPI2 have already been altered by topcoding and anonomyzing. We thought to focus on those variables more closely resembling their original state (that collected by the Census Bureau).

2.1 Predictor Variables

Predictor Variable Table	
Variable	Information
SEX	1 = male, 2 = female
RACEPI2	1 = white only, 2 = Black, 3 = AIAN only, 4 = Asian only, 5 Race group not releasable, 6 = Multiple Race
AGE_P	18-84 = 18-84 years, 85 = 85+ years

The **SEX** variable is a binary variable used to represent gender. The NHIS does not capture the multiplicity of gender identities that members of society can have, instead it is restricted to the gender binary. Because we are using data collected by the government, we will have to conform to their restrictions.

The **RACERPI2** variable is a six-level categorical variable describing race. It is not the only variable pertaining to race in the Sample Adult dataset, however, it is the most up-to-date in terms of OMB standards.

Finally, **AGE_P** is the age variable in the survey. It is integer valued and top coded at 85 years:

2.2 Synthesized Variables

We synthesized the variables FLA1AR, DOINGLWA and WKDAYR in our analysis. All of these variables, like the above, have non-recorded options represented as variables. I will not be considering rows that have such unrecorded options.

Synthesized Variable Table	
Variable	Information
FLA1AR	1 = limited in any way, 2 = not limited in any way
DOINGLWA	1 = Working for pay at a job or business, 2 = With a job or business but not at work, 3 = Looking for work, 4 = Working, but not for pay, at a family-owned job or business, 5 = Not working and not looking for work
WKDAYR	000 = None, 001-366 = 1-366 days

FLA1AR is a binary varibale denoting if a participant has any functional limitation. We chose not to look at whether or not that limitation was chronic since over 95 percent of people experiencing a functional limitation have one such limitation which is chronic. As part of the cleaning process, we delete the rows such where $FLA1AR = 3$.

In the NHIS Sample adult dataset, the **DOINGLWA** is a eight-level categorical variable trying to capture a participants work status in the week before participating in this study. Three of those options are non-recorded, so they are not considered.

Finally, we are interested in the number of lost work days participants needed to take during the past year due to health problems. Only 1.7% of participants in this survey lost more than 50 workdays in the previous year and only 0.6% lost more than 100 work days. Participants under these constraints are particularly vulnerable to being identified as members of this survey.

3 Analysis

3.1 Sequential Synthesis

To create our synthetic dataset, we generate it from the posterior predictive distributions. That is, we take one draw of the posterior parameters and we use the distribution defined by those parameters to generate the synthetic data. For example, in the FLA1AR synthesis model (described below), we consider one set of $\{\beta_1, \beta_2, \ldots, \beta_7\}$ to calculate p_i for each row of predictors. We then pull our synthetic value $\tilde{Y}$ from the Bernoulli distribution defined by p_i.

In order to maintain the relationship between our synthesized variables, we use some of them as predictors. This means that we first fit our model with the un-synthesized, or original predictor variables. Then we generate our synthetic data in a sequential manner. Here * denotes synthesized variables:

$$\begin{aligned} \pi(FLA1AR^* \mid SEX, RACERPI2) \\ \pi(DOINGLWA^* \mid FLA1AR^*, SEX, RACEPI2) \\ \pi(WKDAYR^* \mid AGE\_P, FLA1AR^*, DOINGLWA^*). \end{aligned}$$

We call this process *Sequential Synthesis*. It is very similar to the process of multiple imputation for missing data [4].

3.2 Models

We use three distinct types of models for our synthesized variables depending on the type of outcome each variable has. Note: we consider all outcome observations Y to be independent and identically distributed. For the binary variable **FLA1AR** we use a Bayesian logistic regression with SEX and RACERPI2 as predictors. This model is appropriate as it fits binary outcome and allows us to use categorical predictor variables. We outline the model below:

$$\begin{aligned} Y \sim & \text{Bernoulli}(p) \\ \text{logit}(p) = & \beta_1 + \beta_2 \cdot SEX_{female} + \beta_3 \cdot RACE_{black} + \beta_4 \cdot RACE_{AIAN} + \beta_5 \cdot RACE_{asian} \\ & + \beta_6 \cdot RACE_{not\ releasable} + \beta_7 \cdot RACE_{multiple} \\ \beta_i \sim & \text{Normal}(\mu, \sigma) \end{aligned}$$

In the model we implemented, we let $\mu = 0$ and $\sigma = 0.1$.

We use a similar approach for modeling the **DOINGLWA** variable. Instead of using simple logistic regression, we must use multinomial logistic regression as DOINGLWA has five possible outcomes. This way we can fit all outcomes and give them predictor variables:

$$\begin{aligned} Y \sim & \text{Categorical}(p_c) \\ \log(q_c) \sim & \beta_1 + \beta_2 \cdot FLA1AR_{limited} + \beta_3 \cdot SEX_{female} + \beta_4 \cdot RACE_{black} + \beta_5 \cdot RACE_{AIAN} \\ & + \beta_6 \cdot RACE_{asian} + \beta_7 \cdot RACE_{not\ releasable} + \beta_8 \cdot RACE_{multiple} \\ p_c = & q_c \cdot \frac{1}{\sum_{i=1}^{c} q_c} \\ \beta_i \sim & \text{Normal}(\mu, \sigma) \end{aligned}$$

In the model we implemented, we again let $\mu = 0$ and $\sigma = 0.1$. Here $C = 5$, since it is the number of possible outcomes the DOINGLWA can take. We let $c \in \{1, 2, \ldots, C\}$.

For the count of missed work days in a year, ie. the **WKDAYR** variable, we use Poisson Regression. Similar to the logistic regressions we ran on the previous variables, we can also give that Poisson regression linear predictors. We will use our two to-be-synthesized variables as predictors as well as the AGE_P variable. Note: we also added an error term η to our model. This allows us to partially correct for the overdispersion of our model. In a Poisson model, it is assumed that the rate of variance over mean is one. In our model it is 88.5147. This means that the Poisson model will add noise by more regularly dispersing the count of missed days does the original data[1].

$$\begin{aligned} Y \sim & \text{Poisson}(\lambda) \\ \log(\lambda) = & \beta_0 + \beta_1 \cdot AGE\_P + \beta_2 \cdot FLA1AR_{no\ lim} + \beta_3 \cdot DOINGLWA_{working} \\ & + \beta_4 \cdot DOINGLWA_{vacation} + \beta_5 \cdot DOINGLWA_{looking} \\ & + \beta_6 \cdot DOINGLWA_{work\ not\ for\ pay} + DOINGLWA_{not\ looking} + \eta \\ \beta_i \sim & \text{Normal}(\mu, \sigma) \\ \eta \sim & \text{Normal}(0, \phi) \\ \phi \sim & \text{Gamma}(a, b) \end{aligned}$$

Here, we let $\mu = 0$, $\sigma = 0.1$, $a = 1$ and $b = 1$.

4 Results

Overall, our sequential synthesis approach results with high utility. Figures 1, 2 and 3 compare the distributions of the original data and our synthesized data for the FLA1AR, DOINGLWA and WKDAYR variables

respectively. The logistic and multinomial logistic regressions used for the FLA1AR and DOINGLWA variables worked well to generate a new data set with a similar distribution to the original. The Poisson regression used to synthesize the WKDAYR variable, however, did not capture the skewness of that variable. This is clear from the lack of a linear trend in the scatter plot and the differences in ranges of the original and synthetic data (the original range is 0 to 365 while the synthetic is 0 to 85).

The short-comings of our WKDAYR model appears to be a consequence of the nature of the Poisson model itself. In the Poisson model, it is assumed that the ratio of the variance and the mean is one. In our count data, that ratio is 88. This means that WKDAYR count data is over-dispersed. To account for the over-dispersion, we added an error term η to our linear predictors. However, given time constraints and the computational intensity of our model, we have concluded that the model we used to generate synthetic data displayed in Figure 3 did not converge.

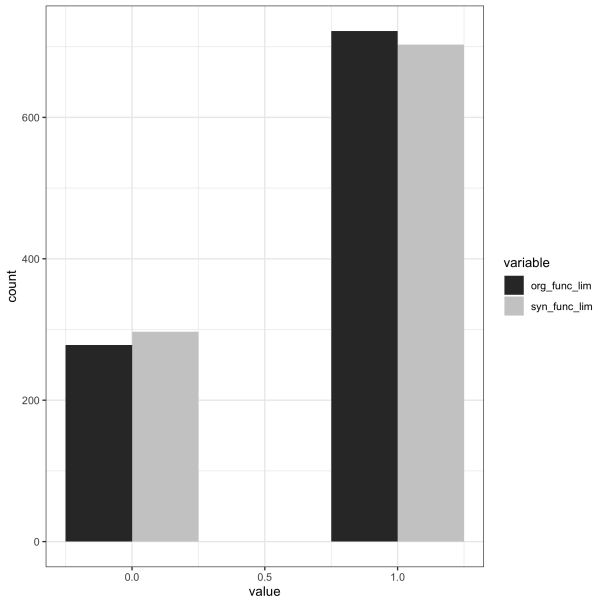

Figure 1: FLA1AR Synthesis

4.1 Utility

We will use two different measures to try to capture the utility of our data. First, we consider the global *Propensity Score Measure*. Given two data sets, one synthesized and other un-synthesized, the *Propensity Score* for an element in either data set is the probability that it is in the synthesized data set. We calculate this measure by, first, merging the original and synthesized data. The merged data contains an indicator variable T which is 0 if the element is from the original data and 1 if the element is synthetic. We can then estimate propensity score by running a logistic regression on T on functions of the synthesized variables. We can then assess the similarity of propensity scores by comparing the percentiles of each group. We can simply compute

$$U_p = \frac{1}{N}\sum_{i=1}^{N}[p_i - c]^2,$$

where N is the total entries in the merged set, p_i is the propensity score for each element and c is the proportion of units from the synthesized data[5]. For our purposes, N is double the number of records in the

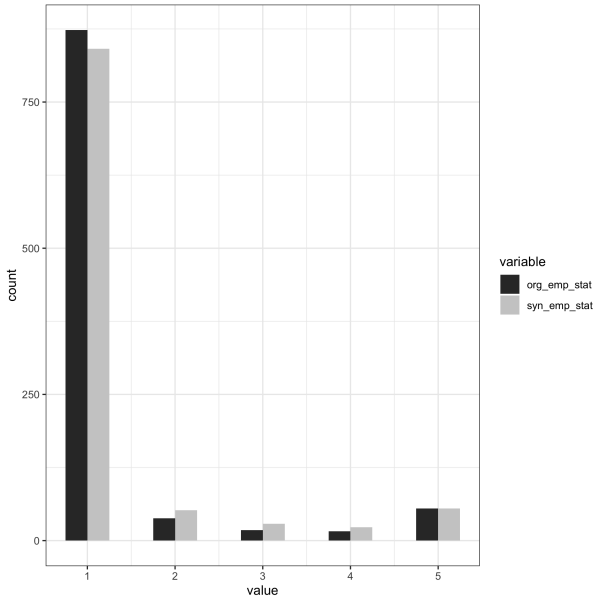

Figure 2: DOINGLWA Synthesis

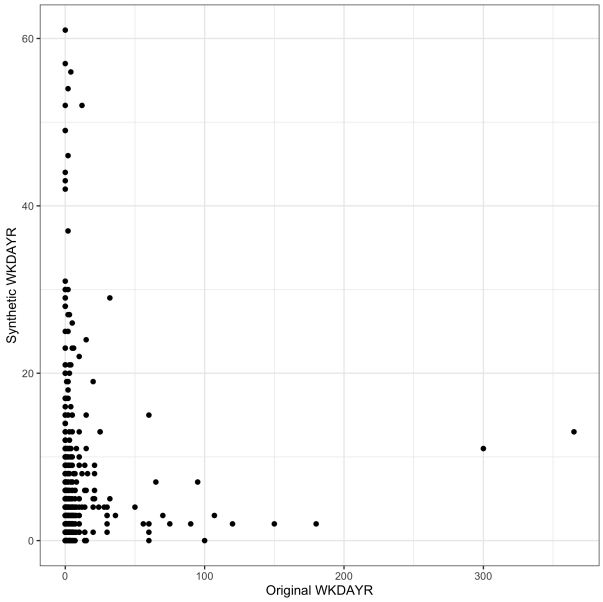

Figure 3: WKDAYR Synthesis

synthetic or set and $c = \frac{1}{2}$.

We can use *Cluster Analysis*, a form of unsupervised machine learning, to preform another global utility measure. First, we again merge the original and synthesized data, where O denotes the original data and S is synthesized. Then, we use Cluster Analysis to place records with similar values of selected variables into

G groups. Then we can calculate the below measure:

$$U_c = \frac{1}{G}\sum_{j=1}^{G} w_j \left[\frac{n_{jO}}{n_j} - c\right]^2,$$

where n_j is the number of elements in the jth cluster, n_{jO} are the number of elements in the jth cluster from the original data, w_j is the weight for the jth cluster and $c = N_O/(N_O + N_S)$[5]. Again, for our purposes $c = \frac{1}{2}$.

Variable	U_p	U_c
FLA1AR	0.0001101	0.05
DOINLWA	$2.2728e^{-5}$	0.04725
WKDAYR	0.0002071	0.0001090

Table 1: Global Utility Measures

All values in the above table are close to zero. This indicates high utility. This high utility contrasts the observed difference in ranges for the original and synthesized distribution of the WKDAYR variable.

4.2 Disclosure Risk

First, we consider the risk for identification disclosure within our data set. We calculate three summary measures of identification disclosure: Expected Match Risk, True Match Rate and False Match Rate. First, consider the values needed to calculate these measures: c_i is a vector indicating for each target record i the number of records with highest match probability. The *Highest Match Probability* is a subset of all records sharing the same (in the case of categorical data) or very similar (in the case of continuous data) original and synthesized variables. Next, T_i is a binary vector equal to one if the true match between the original and synthetic data is within the c_i units, and zero if not. If the true match is unique, ie $c_i \cdot T_i = 1$, then $K_i = 1$, otherwise $K_i = 0$. Conversely, if there is a unique match but it is not the true match, ie $c_i(1 - T_i) = 1$, then $F_i = 1$, otherwise $F_i = 0$. Also, N is the total number of target records, which for our purposes is the total number of records in our NHIS sample, and s is the total number of uniquly matched records (ie. $\sum_{i=1}^{N} c_1 = 1$). We can preform the following calculations:

- the *Expected Match Risk:*

$$\begin{cases} \sum_{i=1}^{N} \frac{T_i}{c_i}, \text{if } c_1 >= 1, \\ 0, \text{if } c_i = 0. \end{cases}$$

  Here if $T_i = 1$, then we are computing the probability for the intruder will pick the true, then the value T_i/c_i is zero.

- The *True Match Risk:*

$$\sum_{i=1}^{n} \frac{K_i}{N},$$

  where n is the number of records for which we wish to calculate this risk score. In our calculations, we let $n = N$.

- the *False Match Rate:*

$$\sum_{i=1}^{N} \frac{F_i}{s}.$$

  This is the percentage of false matches among unique matches[3].

Since the WKDAYR variable is continuous, we consider the highest match probability to contain those records whose original WKDAYR values are between 20 percent above or below the synthetic value. Below is a table of our results.

Expected Match Risk	True Match Rate	False Match Rate
11.7833	0.007	0.8793103

Table 2: Identification Risk Measures

To understand the expected match risk in the context of our data set we can divide it by the total number of records in our data set: $11.7833/1000 = 0.0117833$. Since this value and the true match rate are close to 0 and the false match rate is sufficiently close to 1, we conclude that this synthetic data set has sufficiently small identification disclosure risk.

5 Discussion

We consider our models to be effective in managing the tension between maintaining high utility while accounting for disclosure risk. However, our models clearly have shortcomings. Most obvious is the problem with our WKDAYR Poisson regression. The for our synthetic data is 0 to 61 while the range in the original was 0 to 365. Our utility according to global utility measures remain still quite high. Our predictors for the DOINGLWA synthesis may have been too effective. In an earlier iteration of our model (not using the 1000 variable sample), the addition of linear predictors what was a Dirichlet-Multinomial Model decresed the false match rate from 0.99 to 0.93. We could use fewer or less effective predictors to make sure to add enough noise to increase participant privacy.

In the future, we would like to do the analysis on the entire 25417 observations contained in the 2018 NHIS Sample Adult data set. We would also like to consider synthesizing dependent outcome variables created by a sequence of related questions.

References

[1] Ronald Christensen, Wesley Johnson, Adam Branscum, and Timothy E. Hanson. *Bayesian Ideas and Data Analysis*, chapter 11. CRC Press, Florida, 2011.

[2] National Center for Health Statistics. National health interview survey, 2018. public-use data file and documentation., 2019. Last accessed 6 May 2020.

[3] Jingchen Hu. Bayesian estimation of attribute disclosure and identification disclosure in synthetic data. *Transactions on Data Privacy*, (12):61–89, 2019.

[4] Satkartar K. Kinney, Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, RonS. Jarmin, and John M. Abowd. Towards unrestricted public use microdata: The synthetic longitudinal business database. *International Statistical Review*, 79(3):362–384, 2011.

[5] Mi-Ja Woo, Jerome P. Reiter, Anna Oganian, , and Alan F. Karr. Global measures of utility for microdata masked for disclosure limitation. *The Journal of Privacy and Confidentiality*, 1(1):111–124, 2009.