Prisoners Are People Too: Statistical Disclosure Control in the 2016 Survey of Prison Inmates

Nicole Mack[1], Marcus Berzofsky[1], Stephanie Zimmer[1]
[1]RTI International , 3040 E Cornwallis Road, Durham, NC, 27709

Abstract
Organizations releasing public or restricted use files attempt to minimize disclosure risks for participants while maintaining data utility. This presentation reviews the disclosure risk assessment for the 2016 Survey of Prison Inmates (SPI). SPI routinely collects identifiable and sensitive information such as health histories and offense histories from prisoners. Surveys of confined populations, such as prisoners, pose higher disclosure risk challenges than general population surveys. Populations in known, confined locations - like prisons - are more vulnerable to identification. Thus, we sought to determine the methods that best encapsulate the trade-off between data utility and minimization of disclosure risk. Through a variety of assessments of frequencies and uniqueness we discern how easily one can identify certain groups and individuals. We considered non-perturbative and disclosure avoidance methods such as coarsening and suppression to decrease disclosure risk. For each method we evaluate the trade-off and propose how methods presented here could also be applied to other surveys with hierarchal groups such as schools or hospitals.

Key Words: data confidentiality, data quality, correctional facilities

1. Introduction

When persons or entities consent to taking part of a survey that collects confidential, sensitive, and personally identifiable information there is an understanding and trust that such information will be kept safe; that re-identification of a respondent is unlikely to happen and that their data will be used for statistical purposes only. While restricting data access to certain institutions or individuals can provide an initial layer of security, restricted use files where users may utilize the files for privacy invasion or files made available to the public still pose a significant threat to confidentiality of participants. Therefore, it is imperative that other methods be employed to retain anonymity of participants. The field of statistical disclosure control (SDC) seeks to provide that additional layer of security. SDC can be defined as a collection of methods intended to treat and alter data so that such data can be published or released without revealing confidential information while simultaneously seeking to minimize information loss due to anonymization of data (Benschop and Machinguata, 2016). Whereas anonymizing individual level data can be challenging in and of itself, surveys, such as the focus of this paper the Survey of Prisons Inmates (SPI), which employ a clustered design present an additional challenge. Given information collected at the cluster level, in the case of SPI the clusters are prisons, is the same for all individuals in a cluster, re-identification of one individual in a cluster allows for re-dentification of others. Furthermore due to this hierarchal structure values of variable for others in a cluster that are common for can be used to re-identify other individuals within a cluster (Benschop and Machinguata, 2016).

Regardless of data structure, the process of disclosure always includes disclosure risk, utility, and an established threshold. As Benschop and Machinguata (2016) describe disclosure risk occurs when an intolerable estimation of a respondent's confidential information is possible o exact disclosure is possible with a high level of confidence; utility as a concept describes the value of data as a resource comprising analytical completeness and validity; and an established threshold essentially seeks to strike a balance between both disclosure risk and utility or serve as level or point from either from which a data release is deemed safe or unsafe but also strikes For each possible data release measures of risk and utility a threshold should be defined and tailored for the particular dataset.

Once these three components of the process are defined one must decide which methods of disclosure are best to apply. There is a plethora of methods of which some we will describe here. Methods of disclosure are often classified based on whether the method is perturbative or non-perturbative and whether the method is probabilistic or deterministic. Non-perturbative methods are methods where detail within data is reduced through generalization or masking (Hundepool, Domingo-Ferrer, Franconi. Giessing, Nordholt, Spicer, and P.-P. de Wolf, 2012) Perturbative methods are methods where values are altered to create uncertainty around true values. (Hundepool, Domingo-Ferrer, Franconi. Giessing, Nordholt, Spicer, and P.-P. de Wolf, 2012) Probabilistic methods are based on randomness whereas deterministic methods follow a certain algorithm to produce the same results (Templ, Meindl, and Kowarik, 2018; Hundepool, Domingo-Ferrer, Franconi. Giessing, Nordholt, Spicer, and de Wolf, 2012) These types of methods can be applied to both categorical and continuous variables.

Common non-perturbative methods include recoding, sometimes referred to as coarsening, and variable suppression. Recoding, decreasing the number of distinct categories or values of a variable, can be applied to both categorical and continuous variables, whereas suppression, induction of missingness of a value or values, is usually applied to categorical variables (Benschop and Machinguata, 2016). Common perturbative methods for categorical variables include post-randomization (PRAM); a method that reclassifies values of a categorical variable into another category based on pre-defined transition probabilities (Benschop and Machinguata, 2016). For continuous variables there are several different perturbative methods utilizing aggregation techniques and simulations of values to uncertainty. (Templ, M., Meindl, B., Kowarik, A., 2016)

We assume that for users of this data in particular, response differences among the varying demographic and socioeconomic classes would be of very high interest and importance. Thus altered values of the data may not be of much interest. Therefore this paper will focus on non-perturbative methods.
Given the nature of our dataset, this paper will focus on non-perturbative methods. While we are obligated to anonymize the file in some way, beyond the question of how. we were primarily interested in 1) the effect of these SDC methods alone or in combination on potential key outcomes and 2) how does the inclusion or exclusion of survey design information impact risk and utility.

2. Methods

2.1 Study Data

The Survey of Prison Inmates (SPI) is a survey conducted with the purpose of obtaining nationally representative estimates on prisoners in state and federal prisons. The study has been conducted periodically by the Bureau of Justice Statistics with iterations in 1991, 1997, 2004 and 2016. The 2016 SPI was conducted for BJS by RTI International. The 2016 SPI employ a two-staged, stratified, clustered design. Prisons-stratified by sex, jurisdiction, and geography and selected proportional to size-were selected in the first stage with prisoners selected in the second stage. The final respondent sample size consisted of 24,848 respondents from 364 participating prisons. Aside from information used for stratification purposes no other information was collected at the prison level.

Interviews were conducted with participants 18 and older from January 2016 to October 2016. Respondents answered a series of questions on various topics including demographic characteristics, medical history, criminal history, drug and alcohol use, socioeconomic characteristics, as well as work performed and services utilized while in prison. Prisoners responded to all modules in the survey. The survey was administered in English and Spanish.

2.2 Risk

Before assessing risk, we removed all direct identifiers-names, addresses, birthdates, and social security numbers from the file and created new identification numbers for respondents. Given the prisons served as clusters, each facility was given an identification number. This facility identification number serves as the cluster identifier. Next, we determined the key variables to be used to measure risk. These key identifiers were chosen based on what information may be already be publicly available such as demographic information and variables unique to this survey population and available to the public as well such as controlling offense. These identifiers include:

- Age (categorical)
- Sexual Orientation
- Gender Identity
- Race/Ethnicity
- Marital Status
- Education Level
- Income (continuous)
- Controlling Offense
- Sentence Length

As described in Benschop and Machinguata (2016), for categorical identifiers we summarized risk by way of these three measures: global risk, hierarchal risk, and *k*-anonymity. Global risk, the expected proportion of all individuals in a sample that could be re-identified, essentially takes the mean of all individuals in a sample as shown in (1). Hierarchal risks are defined as the risk that at least member of a cluster is re-identified; essentially one minus the union of individual disclosure risks as displayed in (2). Third, *k*-anonymity is based on pattern of key variables containing at least *k* units in the microdata. A dataset is said to have reached *k*-anonymity when the count of the number of individuals with a sample frequency is lower than a specified *k* as shown in (3). For our purposes we considered anonymity values of 3 and 5.

(1) Global Risk: $\frac{1}{n}\sum_k f_k r_k$, n is sample size, r_k is individual risk of key variables k that i^{th} individual shares, f_k, frequency counts of key variables

(2) Hierarchal Risk: $1 - \prod_{J=1}^{J} 1 - P(Aj)$, j=j^{th} member of cluster, ,P(A_j) is individual disclosure risk of the j^{th} member

(3) k-anonymity: $\sum_i \ I \ (f_k < k), i$ is the i^{th} record, I is the indicator function

2.3 Utility

Some survey items within the data while not confidential could be considered sensitive. It is some of these items we chose as our outcomes due to that nature and due to likely interest for users of the data. These outcomes were treatment for alcohol and drug use and mental health status. Treatment for alcohol and drug use is a binary variable indicating whether a person -currently or in the past- has received treatment for alcohol or drug use. Likewise, mental health status is also a binary outcome that indicates whether a person has been diagnosed in the past or as of late with having a mental illness.

After establishing our outcomes of interest, we proceeded to define our utility measures of interest. Our first measure was the percentage of missing values for each of the key variables due to anonymization. Our second set of utility measures include the information loss measure (IL1) defined as the sum of the absolute distances between corresponding observations in the raw and anonymized datasets for continuous variables and the overall mean. (Benschop and Machinguata, 2016) The last measure we use to assess utility is mean square error MSE. For each of our outcomes of alcohol/drug treatment and mental health produced a logistic regression equation using our key variables as covariates. The MSE was calculated for the conditional probabilities for all covariates.

2.4 Treatment Scenarios

Our interests focused on non-perturbative methods of recoding, suppression, and the combination of the two. We wanted to understand the effect of modulating recoding and/or suppressions to extremes of either technique, i.e. little risk reduction to extreme risk reduction, on data quality. In all modified scenarios income was top coded at a value of \$90,000. For our analysis we considered 5 scenarios – the original data and four treatment scenarios. Scenario 0, the original data, is treated as the "truth" for comparison purposes. Since no information was collected at the cluster level aside from stratifying variables, all disclosure methods were applied at the individual level first and foremost. Risk calculations and utility measurements were computed using R v3.4.4 and SAS v9.3. Table 1 summarizes each of the treatment scenarios enacted upon the data.

Table 1: Treatment Scenarios

Scenario 0	Original Data
Scenario 1	Top coding of Income, recoding Controlling Offense from 22 levels to 5, Global Suppression of Facility Identifier (Cluster variable)
Scenario 2	Top coding of Income and local suppression at *k*=5 level, Facility Identifier included

Scenario 3	Top coding Income, recoding Controlling Offense from 22 levels to 5,Facility Identifier included, and local suppression at *k=3*
Scenario 4	Top coding income, recoding controlling offense from 22 levels to 5, local suppression at *k=3*, and Global suppression of Facility Identifier

3. Results

3.1 Risk Assessment

Table 2 shows risk measures for each of the scenarios. Baseline global risk and hierarchal risk are 5.0% and 74.9%, respectively, with the global risk indicating 1,242 expected re-identifications. The least riskiest scenario appears for scenario 2 where local suppression at *k*=5 reduces global risk to 0.1%. The risk of re-identification of a cluster is substantially reduced from baseline to 6.1%. Along this continuum of suppression, the first scenario where no suppression is induced results in a slight decrease in the global risk to 3.3; violations of 3- and 5- anonymity are moderately reduced. In scenarios 3 and 4 where there is a mix of recoding and suppression the risks are significantly reduced from baseline data in terms of global risk, 0.2% and 0.4% respectively. Furthermore, where information on cluster structure was kept the hierarchal risk significantly declined as well with a 62% change. The dash within the table represents instances where the cluster variable was globally suppressed.

Table 2: Risk and Utility results per scenario

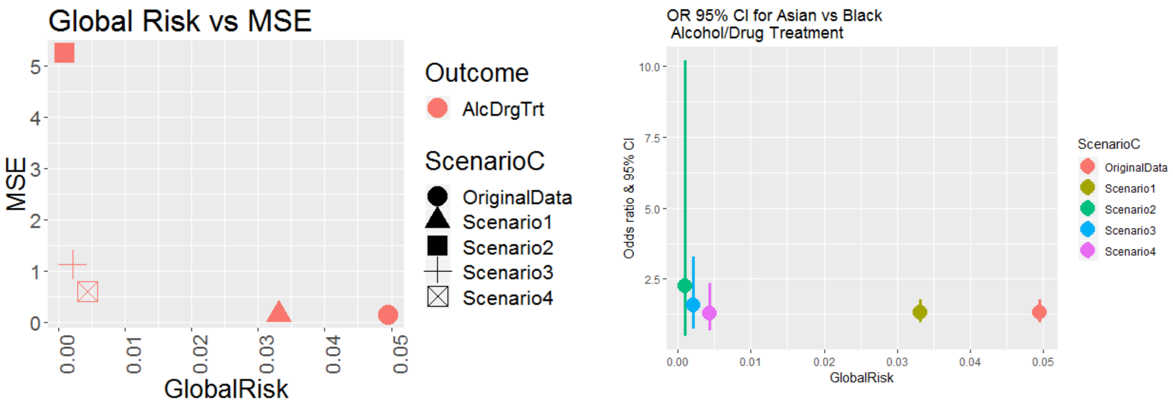
Risk Measures	**Scenario 0**	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**
Global: % of Re-identifications	5.0%	3.3%	0.1%	0.2%	0.4%
Hierarchal Risk	74.9%	-	6.1%	12.9%	-
% Violating 3-anonimity	41.2%	26.3%	0.0%	0.0%	0.0%
% Violating 5-anonimity	55.6%	39.2%	0.0%	1.2%	9.5%
Mean; IL1s (Income)	3771.01; 0	2393.8; 0.108	2393.8; 0.108	2393.8; 0.108	2393.8; 0.108
% Missing of Key Variables due to anonymization	0.0%	0.0%	55.5%	41.4%	25.3%

3.2 Utility Assessment

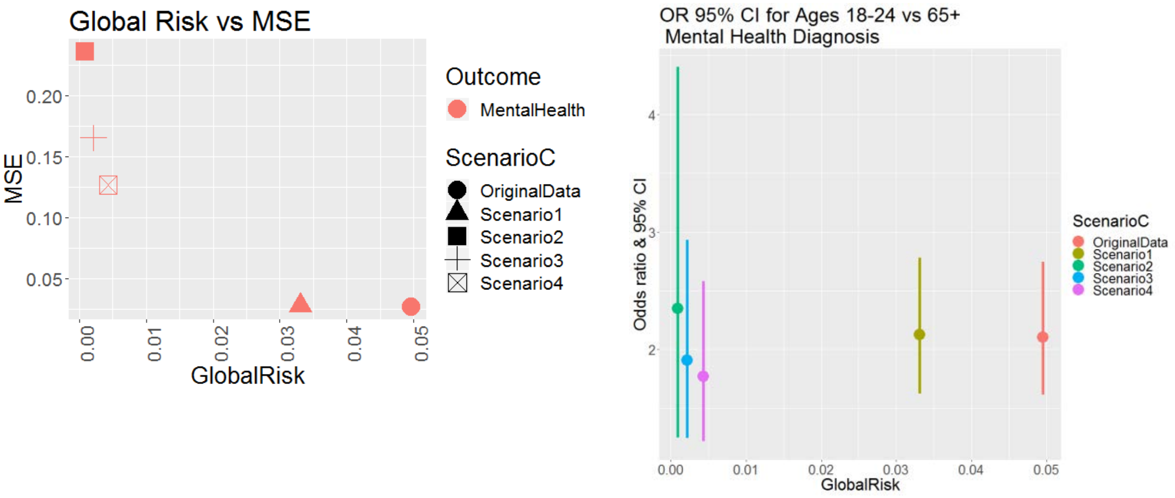
Accordingly, where there was significant reduction in risk there was also a greater percentage of missingness of key variables. As shown in Table 2, for the scenario with the highest level of suppression 55.5% (scenario 2) of key variables are missing. Where there was a mixture of lower levels of suppression and recoding there was substantial but less amount of missingness at 41.4% (scenario 3) and 25.3% (scenario 4), respectively. Although within scenario 3 there was the same level of suppression applied as scenario 4, *k*=3, the rate of missing for scenario 4 is less than scenario 3 despite the risk of scenario 3 being lower. This indicates that although the suppression level and recoding is the same, when one includes information on hierarchal risk, the original risk is therefore higher and therefore suppression of values increase to reach that level of anonymity. Once top coded the information loss for income is minimal at 0.108.

Figures 1a and 1b display the effect of each scenario on mean square error as well as odds ratios from the logistic model of our first outcome. In comparison to the original data, minimal recoding and extreme suppression of scenario 2 results in a high lack of precision (MSE=5.25) and an extremely wide confidence interval (OR=2.247; 0.496, 10.182) for an already small population of Asian prisoners in comparison to Black prisoners. Other scenarios give a much narrower confidence intervals and MSEs with scenario 1 providing the smallest MSE (MSE=0.153) value and odds ratio confidence interval (OR=1.311; 0.967, 1.788) as compared to the original data.

Figures 2a and 2b show a similar pattern to the previous despite the outcome now being mental health status and the covariates in question being comparable in sizes (Age 18-24 vs 65+). Among the treated data the MSE and odds ratios are best for scenario 1 (MSE=0.028, 2.126 (1.625,2.781)). As expected scenario 4 (MSE=0.165, 1.773 (1.1219 ,2.578)) slightly outperforms scenario 3 (MSE=0.236, 2.349 (1.253,4.401)) in terms of smaller MSE values and narrower confidence intervals.



Figures 1a and 1b: MSE and Odds Ratio for Asian vs Black , for Alcohol and Drug Treatment per Scenario



Figures 2a and 2b: MSE and Odds Ratio for Ages 60-65 vs Ages 18-24 , for Mental Health Diagnosis per Scenario

4. Discussion

Statistical disclosure is a complex process highly dependent on assumptions on data users and their intent as well as what is tolerable in terms of risk and utility. For our assessment, we determined that a risk threshold where there are less than 1,000 participants at risk is acceptable. This threshold also comes with the condition of minimizing the amount of missingness as much as possible. As it regards including the cluster variable we recognized that given this risk will always be higher than a dataset that does not include such information, there will take a higher number of suppressions to decrease that risk. For our data this in turn lead to a less riskier data set coupled with a higher degree of missingness- given all other recoding and levels of suppression remain the same. We see this in scenario 3 vs scenario 4. Therefore we decided to not include the cluster variable. For similar surveys with other unique, confined, clustered populations such as hospitals, schools, or military units a similar case could be made for excluding such a design variable as well. Ultimately, we decided not to add to the amount of missingness by inducing local suppressions. Based on all of these observations, we chose to use data transformed in scenario 1. The risk measures chosen were chosen to give a broad and detailed assessment of the individual as well as cluster risks of the data. The utilities chosen, missingness and MSE in particular, are vital summary components indicating the strength of a dataset and precision and accuracy of estimates. We chose to apply non-perturbative methods to the data for various reasons. One reason includes lack of precedent as previous iterations of the survey did not consistently apply such methods. Moreover in today’s climate it is very imperative to have data that accurately reflects the population it was collected from; especially in this case of such a unique, classed, and vulnerable population. This paper serves as a first step in identifying a way that renders the data still useful for public users seeking information on our current prison populations.

5. Acknowledgement

The authors would like to thank and acknowledge the Bureau of Justice Statistics (BJS) for the use of the 2016 Survey of Prison Inmates data and comments provided by Lauren Glaze. However, we would like to note that the views expressed in this paper are those of the authors only and do not reflect the views or position of BJS or the Department of Justice.

References

Benschop, T. , Machinguata, C., Welch, M. (2016, July) *Statistical Disclosure Control for Microdata: A Practice Guide (Version 1.1)* . Retrieved from: http://www.ihsn.org/projects/sdc-practice

Templ, M., Meindl, B., Kowarik, A. (2018, May 16*). Introduction to Statistical Disclosure Control (SDC)*. Retrieved from: https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E.S. Nordholt, K. Spicer, and P.-P. de Wolf (2012) *Statistical Disclosure Control*. Retrieved from: http://neon.vb.cbs.nl/casc/handbook.htm