

Issues in Estimating Reidentification Risk Using Log-Linear Models in Complex Survey Samples

Lin Li*, Jianzhu Li** and Tom Krenzke***

* Westat, 7501 Wisconsin Avenue, Bethesda, MD 20814, USA

linli@westat.com

** FINRA, 1735 K St NW, Washington, DC 20006, USA

jianzhulee@hotmail.com

*** Westat, 7501 Wisconsin Avenue, Bethesda, MD 20814, USA

tomkrenzke@westat.com

Received 31 January 2024; received in revised form 5 February 2025 and 8 December 2025; accepted 28 December 2025

Abstract. In this paper, we discuss some practical issues encountered when estimating record-level and file-level disclosure risk measures of re-identification in survey microdata under complex survey designs. We use the probabilistic modelling approach based on the Poisson Distribution and log-linear modelling proposed in Skinner and Shlomo (2008) to estimate disclosure risk in survey microdata files. We examine the robustness of their GOF criteria to violations of model assumptions, particularly in the context of complex survey designs and differential survey weights, using a case study and simulations. We also provide guidance for variable selection with insights on how to proceed with the disclosure risk assessment and provide meaningful results. For the case study, we use the complex survey dataset from the Survey of Doctorate Recipients conducted by the National Center for Science and Engineering Statistics. The results of evaluating the disclosure risk estimates under different approaches of adjusting the probabilistic modelling to account for the complex survey data lead to guidance for a sensitivity analysis that helps to provide better estimates of record-level and file-level risk of re-identification in survey microdata.

1 Introduction

Statistical agencies are obligated to protect respondents' identities when they release survey microdata to the public. The microdata from the sample survey

often go through statistical confidentiality treatment (e.g., recoding) before being released. To determine whether the treated data is safe to be released, agencies need to assess the disclosure risk of the sample microdata. There are three main types of disclosure risks: identification, attribute, and inferential.

Identity disclosure is about learning a person's name or an establishment's name given the data product. The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) enacted as Title V of the E-Government Act of 2002 is an example of a Federal law that protects against identity disclosure. In general, an identity disclosure takes place when a correct record re-identification is achieved by an intruder (denoted as a person or agency with malicious intent to reveal information) by comparing a target individual in the sample microdata using a set of quasi-identifiers with an available list of individuals in the population that contains personal identifiers such as name and address (see Willenborg and de Waal, 2001). Duncan, Elliot, and Salazar-González (2011) define re-identification as "The process of determining the identity of data subjects in data releases that have been subject to statistical disclosure limitation, that is, have been de-identified."

Individual attribute disclosure can occur after a re-identification has been made and the intruder has access to all the sensitive survey variables and can attribute them to the individual. Group Attribute disclosure does not necessarily lead to revealing an individual's private information, but we are able to learn sensitive information about a group of data subjects, for example, that all individuals in a locality collect benefits.

Inferential disclosure goes beyond group attribute disclosure in that the intruder can infer knowledge about a data subject with high probability through manipulating data releases. Whilst inferential disclosure is not part of the traditional disclosure risk scenarios, it is now gaining in importance due to the proliferation of government and non-government data that is made available to the public via the internet. There are particular concerns that microdata can be reconstructed from multiple data releases from a single data source, such as the Census, and hence there have been recent developments in applying more formal privacy guarantees, for example the Differential Privacy approach (Abowd et al. 2019, and Dwork et al. 2006).

In this paper, we are focusing on the risk of re-identification which can lead to individual attribute disclosure in survey microdata. Even when data have been protected using more formal privacy guarantees to protect against reconstruction, such as Differential Privacy, there is still the need to check for

potential re-identifications. Moreover, re-identification risk should be quantified because the laws (e.g., CIPSEA) are written with re-identification risk in mind.

Hundepool et al. (2012) discuss the context that a re-identification in survey microdata is achieved by an intruder when comparing a target individual in a sample with an available list of units (external file) that contains individual identifiers (e.g., name and address), plus a set of quasi-identifying variables, such as age, sex, race and location. Therefore, it follows that an ideal situation for quantifying the disclosure risk of releasing survey microdata is when a full list of the target population exists (or is accessible) with variables that have the exact same definition existing in both the survey microdata and the target population list. Re-identification occurs when a unit in the released file and a unit in the external file are linked and belong to the same individual in the population. If a population file is available but does not have the exact same variable definitions as the survey microdata, or the data sources suffer from errors, a probabilistic record linkage can be carried out (Fellegi and Sunter 1969).

In the sample microdata file, many cases have a single sample count (referred to as “sample uniques”) by cross-classifying a relatively small number of quasi-identifying (categorical) variables having specific values (referred to as “key variables” hereafter) that may exist in external files. If we take the subset of those records in the microdata that are unique in the cells of the cross-classified key variables, an exact matching process can be conducted by matching the sample uniques to the target population file using the common variables as the matching key. Then the match probability M_k for each cell k is: f_k/F_k where f_k is the sample size in cell k (and in this case, $f_k = 1$) and F_k is the number of corresponding individuals in the population file. If each sample unique in the survey microdata file matched to only one population file record, the risk is 1 (or 100%) for that record. If the number of records that matched to a sample unique is 100, then the record-level risk is 0.01 (or 1%). For each sample unique, we can aggregate the match probabilities to obtain the file-level disclosure risk measure defined as the expected number of correct matches. Another file-level disclosure risk measure of interest is the number of sample unique cells $f_k = 1$ in the cross-classified quasi-identifying key variables that are also population uniques: $F_k = 1$.

Often the population contingency table is not directly available and too expensive to obtain, hence the need for estimating the population frequencies from the sample counts to estimate the disclosure risk.

This article focuses on estimating the re-identification risk when a target population file is not available and therefore we need to use probabilistic modelling to estimate the disclosure risk measures (see examples of probabilistic modeling: Bethlehem, Keller, and Pannekoek 1990; Rinott 2003; Poletini 2003; Reiter 2005). More specifically, we focus on the log-linear modelling approach under the Poisson Distribution as developed by Skinner and Holmes (1998), Elamir and Skinner (2006) and Skinner and Shlomo (2008).

We implemented the log-linear modeling approach to estimate disclosure risk in a number of survey microdata files for U.S. government agencies and internationally. In Skinner and Shlomo (2008), the authors developed Goodness-of-Fit (GOF) criteria to select the appropriate model for the log-linear modeling which minimizes the bias of the estimated disclosure risk measures. Several challenges emerge that relate to satisfying GOF criteria of the log-linear models in the presence of model assumption violations and handling large numbers of variables. The purpose of this paper is to discuss the practical issues encountered and investigate ways to address the challenges. We seek to answer the following research questions:

- What are the likely violations of the log-linear modeling assumptions when implementing in sample surveys with complex designs?
- For common violations of assumptions, is the GOF statistic still a good indicator of disclosure risk estimates being too high or too low? If not, what is the impact on the bias associated with the risk estimates? How can the bias be reduced?
- What guidance can be given to achieve a better estimate of the overall file-level re-identification risk in survey microdata?

We first review the theoretical background of the log-linear modeling under the Poisson Distribution approach as set out in Skinner and Shlomo (2008) in Section 2 including an explanation of the GOF criteria for the model selection. In Section 3, we illustrate the practical challenges encountered when carrying out the log-linear modeling under a complex survey design (e.g., stratification) and differential survey weights. We also introduce the case study data: the 2017 Survey of Doctorate Recipients (SDR) public use file (PUF). The 2017 SDR has a complex sample design that violates the log-linear model assumptions on both independent selection and equal inclusion probabilities within cells. In addition, the strata variables are not available on the PUF to inform the estimation in log-linear modeling. We observed unusual behavior of GOF for the 2017 SDR PUF. Based on the case study data, we describe in Section 4 a series of simulations

motivated by these challenges. The simulations are designed to explore the accuracy of estimating disclosure risk measures based on the GOF criteria. In Section 5, we return to an application using the 2019 cohort of Survey of Doctorate Recipients (SDR) restricted use file based on the findings from the simulations in Section 4. Lastly, conclusions are provided in Section 6 that help to answer the research questions.

2 Approach To Quantify Re-Identification Risk

Before reviewing the log-linear modeling approach for estimating the risk of reidentification in detail, the following cautionary notes are provided to accompany the results of the disclosure risk assessment:

- It is assumed that the intruder has no response knowledge on who is included in the survey sample dataset.
- The risk assessment only addresses the re-identification risk, not attribute disclosure or inferential disclosure.
- The disclosure risk measures derived from modeling rely heavily on assumptions made on the amount of information available to intruders. For example, if variables are recoded, then the modelling reflects intruders' less detailed knowledge of respondents' characteristics.
- The estimation of the re-identification risk focuses on the sample uniques that are also population uniques, with sample uniques and population uniques defined by the key variables used in the risk assessment.
- The disclosure risk measures do not investigate the likelihood of attack by intruders. They assume the likelihood of attack is 100%.
- The disclosure risk measures assume that an intruder has a certain amount of public information that is common to the survey variables, which allows him/her to successfully identify some data subjects in the data file. It does not account for the situation that intruders with different sources of public information may form a coalition.

In addition, as recognized by Skinner and Shlomo (2008) the log-linear modeling approach has known limitations under certain circumstances, which can render its application unfeasible. When the adjustments proposed by them

cannot be implemented—particularly due to the unavailability of strata variables—the procedure should not be used.

The probabilistic modeling framework developed by Skinner and Holmes (1998) and Elamir and Skinner (2006) takes a simplified approach that restricts the information that would be known to intruders and does not assume possible attack scenarios. A summary of their approach is described in Shlomo and Skinner (2022) and we mention the approach briefly here to ensure clarity in this paper:

We denote F_k the population size in cell k of a table spanned by key variables having K cells, f_k the sample size in cell k , $\sum_k F_k = N$ and $\sum_k f_k = n$. The set of sample uniques, is defined: $SU = \{k: f_k = 1\}$ and these are the high-risk records with the potential to be population uniques. Two global disclosure risk measures (where I is the indicator function) are the following:

1. Number of sample uniques that are population uniques:

$$\tau_1 = \sum_k I(f_k = 1, F_k = 1)$$
2. Expected number of correct links if we were to match the sample uniques to the population (assuming a random assignment of the population within cell k): $\tau_2 = \sum_k I(f_k = 1) 1 / F_k$.

It is generally assumed that the population frequencies F_k are unknown and we use probabilistic modelling to estimate the disclosure risk measures as follows:

$$\begin{aligned} \hat{\tau}_1 &= \sum_k I(f_k = 1) \hat{P}(F_k = 1 | f_k = 1) \text{ and} \\ \hat{\tau}_2 &= \sum_k I(f_k = 1) \hat{E}(1 / F_k | f_k = 1). \end{aligned} \quad (1)$$

Given that we are modelling population counts based on a contingency table of sample counts spanned by the key variables, we assume a Poisson distribution and a log-linear model to estimate disclosure risk measures in (1). In this model, F_k are realizations of independent Poisson random variables: $F_k \sim Pois(\lambda_k)$ for each cell k . A sample is drawn by Poisson or Bernoulli sampling with a sampling fraction π_k in cell k : $f_k | F_k \sim Bin(F_k, \pi_k)$. It follows that:

$$f_k \sim Pois(\pi_k \lambda_k) \text{ and } F_k | f_k \sim Pois(\lambda_k (1 - \pi_k)) \quad (2)$$

and population cell counts F_k given the sample cell counts f_k are also realizations of independent Poisson random variables.

As typical in this type of framework, the parameters λ_k are estimated using log-linear modeling. The sample frequencies f_k are independent Poisson distributed with a mean of $\mu_k = \pi_k \lambda_k$. A log-linear model for the μ_k is expressed as: $\log(\mu_k) = \mathbf{x}_k' \boldsymbol{\beta}$ where \mathbf{x}_k is a design vector which denotes the main effects and interactions of the model for the key variables. The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ are obtained by solving the score equations:

$$\sum_k (f_k - \pi_k \exp(\mathbf{x}'_k \boldsymbol{\beta})) \mathbf{x}_k = 0. \quad (3)$$

The fitted values are then calculated by: $\hat{\mu}_k = \exp(\mathbf{x}'_k \hat{\boldsymbol{\beta}})$ and $\hat{\lambda}_k = \hat{\mu}_k / \pi_k$. Individual disclosure risk measures for cell k under the assumption of the Poisson distribution are:

$$\begin{aligned} P(F_k = 1 | f_k = 1) &= \exp(\lambda_k(1 - \pi_k)) \text{ and} \\ E(1/F_k | f_k = 1) &= (1 - \exp(\lambda_k(1 - \pi_k))) / (\lambda_k(1 - \pi_k)). \end{aligned} \quad (4)$$

Plugging $\hat{\lambda}_k$ for λ_k in (4) leads to the record-level disclosure risk measure estimates $\hat{P}(F_k = 1 | f_k = 1)$ and $\hat{E}(1/F_k | f_k = 1)$ and these are aggregated to obtain global disclosure risk measure estimates $\hat{\tau}_1$ and $\hat{\tau}_2$ for (1).

Skinner and Shlomo (2008) developed a method for selecting the main effects and interaction terms for the log-linear model that finds the right balance in accounting for random and structural zeros in the contingency table. The method is based on estimating and (approximately) minimizing the bias of the risk estimates $\hat{\tau}_1$ and $\hat{\tau}_2$. Defining $h(\lambda_k) = P(F_k = 1 | f_k = 1)$ for τ_1 and $h(\lambda_k) = E(1/F_k | f_k = 1)$ for τ_2 , they obtain the following expression:

$$\hat{B} \approx \sum_k \hat{\lambda}_k \exp(-\pi_k \hat{\lambda}_k) (-h'(\lambda_k)(f_k - \pi_k \hat{\lambda}_k) + h''(\lambda_k)((f_k - \pi_k \hat{\lambda}_k)^2 - f_k) / (2\pi_k)). \quad (5)$$

For example, for τ_1 :

$$\begin{aligned} \hat{B}_1 \approx \sum_k \hat{\lambda}_k \exp(-\hat{\lambda}_k) (1 - \pi_k) \{ (f_k - \pi_k \hat{\lambda}_k) + \\ (1 - \pi_k)[(f_k - \pi_k \hat{\lambda}_k)^2 - f_k] / (2\pi_k) \}. \end{aligned} \quad (6)$$

As can be seen, the goodness-of-fit criteria \hat{B} are related to the notion of measuring over and under-dispersion as was developed in the Econometrics literature (Cameron and Trivedi 1998). The method selects the model using a forward search algorithm which minimizes the standardized bias estimate $\hat{B}_i / \sqrt{\hat{v}_i}$ for $\hat{\tau}_i, i = 1, 2$, where \hat{v}_i are variance estimates of \hat{B}_i . The goodness-of-fit criteria $\hat{B}_i / \sqrt{\hat{v}_i}$ have an approximate standard normal distribution (critical value ± 1.96) under the hypothesis that the expected value of \hat{B}_i is zero. In our paper, we focus on the risk measure τ_1 and the GOF statistic expressed as $\hat{B}_1 / \sqrt{\hat{V}_R}$ where we use a robust estimator for the variance of \hat{B}_1 by plugging in estimated $\hat{\lambda}_k$ for λ_k and assuming the f_k are independent with mean and variance equal to

$$\mu_k: V_R = \sum_k \{ a_k (f_k - \mu_k) + b_k [(f_k - \mu_k)^2 - f_k] \}^2,$$

where

$$a_k = (1 - \pi_k) \lambda_k \exp(-\lambda_k) \text{ and } b_k = (1 - \pi_k)^2 \lambda_k \exp(-\lambda_k) / (2\pi_k).$$

The GOF statistic tends to decrease as more interaction terms are added to the model. If the model is under-fit, GOF tends to be positive and the risk is overestimated. In many empirical experiments shown in Skinner and Shlomo

(2008), they found that the independence log-linear model tends to under-fit and leads to overestimation of disclosure risk measures, and the all-three-way interaction model tends to over-fit and leads to underestimation of disclosure risk measures. The authors found that the all two-way interaction log-linear model often leads to good estimates of the risk measures when used in the context of large-scale government surveys drawn randomly from the general population.

The sample rate π_k in each cell k defined by the cross-classified key variables is assumed known. For simple random sampling without replacement and other equal probability sampling designs, you can use a single $\hat{\pi} = n/N$ where n is the sample size and N is the population size. Under complex surveys, we need to estimate π_k in practice. One possible estimator of π_k is the overall sampling rate: $\hat{\pi} = \sum_k f_k / \sum_k \sum_i w_{ki}$ where w_{ki} is the sampling weight for case i in cell k . Another possible estimator of π_k is the sampling rate in each cell k : $\hat{\pi}_k = f_k / \sum_i w_{ki}$.

Social surveys often use complex designs where samples are selected with unequal probabilities. Our focus is on situations where the key assumption, namely, $P(f_k = 1 | F_k) = F_k \pi_k (1 - \pi_k)^{F_k - 1}$, does not hold, due to complex sampling schemes. Shlomo and Skinner (2008) discuss this issue in detail, providing guidance on when the assumption is approximately valid and suggesting adaptations to the procedure in such cases. The goal of this paper is to illustrate how severe departures from the basic assumptions impact the GOF statistic.

For complex designs, the log-linear models need to be fit on weighted sample counts in order to obtain unbiased estimates of population parameters. Skinner and Shlomo (2008) note that survey weights need to be used for complex designs in order to obtain a consistent estimate for λ_k through pseudo-maximum likelihood estimation (Rao and Thomas, 2003), where the estimating equation in (3) is modified as:

$$\sum_k (\hat{F}_k - \exp(\mathbf{x}'_k \boldsymbol{\beta})) \mathbf{x}_k = 0 \quad (7)$$

and \hat{F}_k is obtained by summing the survey weights in cell k : $\hat{F}_k = \sum_{i \in k} w_i$. The resulting estimates λ_k are plugged into expressions in (4) and π_k is replaced by the estimate $\hat{\pi}_k = f_k / \hat{F}_k$. The GOF \hat{B} is also adapted to the pseudo-maximum likelihood method. Skinner and Shlomo (2008) also noted that if the only variation in weights is between strata, then λ_k can be consistently estimated by simply including the strata as a model covariate and including the main effect of this covariate in the log-linear model.

We want to note one of the limitations of log-linear modelling, i.e., the existence of MLE is generally not guaranteed in log-linear models for very sparse tables. As discussed in the general theory (Fienberg and Rinaldo, 2012) and in the context of disclosure risk estimation (Manrique-Vallier and Reiter, 2014; Carota et al., 2022), for sparse contingency tables, log-linear modelling may incur severe issues, one of which is the non-existence of MLE. Notably, fitting routines usually do not check conditions for the existence of MLEs. Nevertheless, we used the Bishop-Fienberg (Bishop, Fienberg, and Holland, 2007) method of iterative proportional fitting (IPF) and also kept the average cell size above 0.01, which is a rule of thumb based on empirical experience to ensure that the IPF will converge. Therefore, we were able to find solutions.

In the following sections, we discuss some practical issues that we encountered under the probabilistic modelling approach of the Poisson Distribution and log-linear modelling through a series of simulations based on a case study of the 2017 Survey of Doctoral Recipients (2017 SDR) public use data.

3 Case Study: 2017 Survey Of Doctoral Recipients

For the case study, we first conducted a re-identification risk assessment on the 2017 Survey of Doctoral Recipients (SDR) Public Use File (PUF) (NCSES 2017). The SDR is a longitudinal survey conducted approximately every 2 years that is designed to provide demographic and career history information about individuals who earned a research doctoral degree in a science, engineering, or health (SEH) field from a U.S. academic institution. The SDR is drawn from the target population of all SEH Doctoral Graduates as the sampling frame. This information is available in the Survey of Earned Doctorates (SED), an annual census conducted since 1957 of all individuals receiving a research doctorate from an accredited U.S. institution in a given academic year. The SDR follows a sample of individuals with SEH doctorates throughout their careers from the year of their degree award until age 76. The panel is refreshed each survey cycle with a sample of new SEH doctoral degree earners (referred to as cohort hereafter). For the 2017 SDR, all 2015 sample members who remained age eligible for the survey were retained, and a sample of new graduates was added. The new graduates sample was selected using a stratified sample design, where the strata were defined by over 200 fine fields of study. The resulting 2017 SDR sample consists of 124,580

individuals with 85,739 respondents. The overall sampling rate from the frame of SEH Doctoral Graduates was about 11% which is very high compared to standard general population surveys, while sampling rates varied greatly across strata. Consequently, the sampling weights have a large variation with a coefficient of variation of 109% (minimum weight: 1 and maximum weight: 89). We note the risk assessment in this case study dataset treated the dataset as cross-sectional and did not account for the disclosure risk arising from the longitudinal nature of the data.

3.1 Disclosure Scenario and Variable Selection

The first step in estimating the risk of re-identification is to determine the key variables to be included in the log-linear models. There were many indirect identifying variables available in the 2017 SDR PUF. Since including too many variables would overstate intruder knowledge, we reduce the number of variables by choosing one variable to represent a group of similar variables. For example, we picked one out of the following variables: year of highest degree, year of most recent degree, and year of first degree. Also, we chose indirect identifiers that can be relatively easily obtained by intruders from external sources over those that were difficult to obtain. Eight variables were included in the log-linear model as shown in Table 1. The average cell size (including the zero sample cells) is 1.44 after cross-classifying eight key variables.

Table 1: Indirect Identifying Variables Used in Re-Identification Risk Assessment for the 2017 SDR PUF

Variable	Number of categories
Academic position	3 (Dean or President; others)
Age group	10 (5-year intervals, bottom coded)
Place of birth	2 (U.S./Non-U.S.)
Total number of children	3 (No child/1 child/2 or more children)
Gender	2 (Male or Female)
Federal government support	3 (Yes, No, Not applicable)
Year of award of highest degree	11 (5-year intervals, bottom coded)
Race/ethnicity	5 (Asian, Black, Hispanic, White, other)

3.2 Model Fitting

We fit the log-linear models with the eight key variables shown in Table 1. To fit the models, we use iterative proportional fitting on a set of margins where each original cell value receives a value of 1 (Bishop, Fienberg, and Holland, 2007). An unusually high level of interaction terms was needed to satisfy the GOF criteria of the log-linear models for the 2017 SDR PUF. As illustrated in Figure 1, the GOF statistic was over 5 even with all five-way interactions in the log-linear model. The poor model fit may be due to violations of model assumptions. Both the 2015 SDR sample and 2017 SDR sample had deep stratification, and sampling rates varied greatly across strata. If the design strata can be included in the key variables, it would help mitigate the violations. However, the strata variable is not available in the PUF rather it appears in collapsed form. Although the agency releasing the PUF has the strata data, we are limited to the variables available in the PUF to carry out the disclosure risk assessment based on assumptions of what the intruder will know. The intruders would only have available those key variables and other variables that are available in the released PUF. The within-cell selection rate π_k is also unknown and difficult to estimate due to the large variation in sampling weights. This raises a question: “Is the GOF statistic still a good indicator of unbiased risk estimates when model assumptions do not hold?”. In Section 4, we present a series of simulations to answer this question.

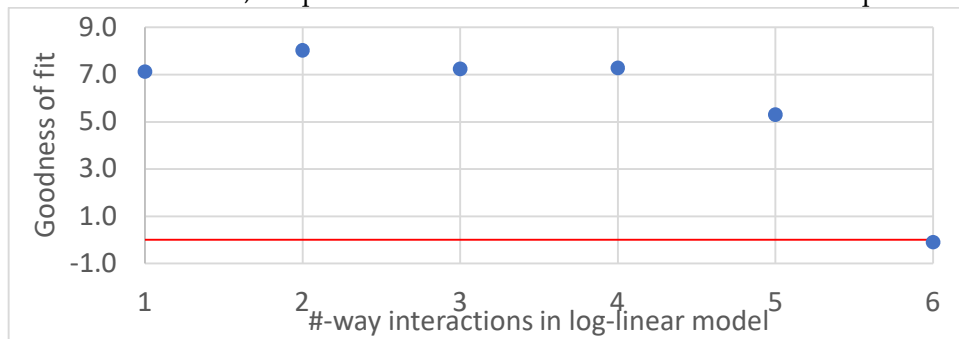


Figure 1. The goodness-of-fit statistic by the level of interaction terms in the log-linear model for the 2017 SDR PUF

4 Simulation Study

The purpose of the simulation study is to explore the accuracy of disclosure risk estimates based on the GOF criteria particularly for the case of complex survey designs and differential survey weights. We do not actually pursue the model search strategy proposed by Skinner and Shlomo (2008) since the model search is not the topic of the paper. Instead, we focus on the issue of volatile survey weights and parts of the sample that have a weight of 1 or 2 and therefore do not meet the assumptions of the original approach. We describe the simulation settings and discuss the results for the nine simulation scenarios in turn below.

4.1 Simulation Setup

We use the 2017 SDR PUF as the population and draw a 1% sample (sample size = 840) 1000 times by simple random sampling (SRS) and stratified sampling, respectively. The SRS sample aligned well with the log-linear model assumptions and was included for verification purposes. The stratified sample was selected in two different fashions: one defined strata using two of the key variables, the other defined strata randomly without using any key variables. Sampling weights were calculated for each sample as the inverse of the selection probability. The sampling weights ranged from about 2 to over 300 with a coefficient of variation of roughly 100% for each of the samples. We fit log-linear models with the counts in cells formed by cross-classifying eight key variables as defined in Table 1 except that the race/ethnicity variable had the categories Black and Other combined to reduce the number of cells. The average cell size is 0.02. The log-linear models were fit in two ways: using sample counts or weighted sample counts. The fitted models provided estimates of the parameters λ_k . We also need to estimate π_k to obtain the disclosure risk estimates and GOF. Two approaches were used to estimate π_k : (1) the estimated overall sampling rate (referred to as $\hat{\pi}$ hereafter), and (2) the estimated cell sampling rate (referred to as $\hat{\pi}_k$ hereafter). The outcome measure for the simulation is the difference between the risk estimated from the fitted log-linear model and the true risk based on the simulation sample and population (which is the 2017 SDR PUF). A positive difference indicates overestimation of risk, and vice versa. For risk estimates we used τ_1 which is the number of sample uniques that are also population uniques.

After running through the simulation, we examined the relative difference of risk estimates against the GOF statistic, where the relative difference was defined as $RD = (estimated \tau_1 - true \tau_1) / true \tau_1$. For the GOF statistic, we used the $\hat{B}_1 / \sqrt{V_R}$ proposed in Skinner and Shlomo (2008). In total, there are nine simulation scenarios as illustrated in Figure 2.

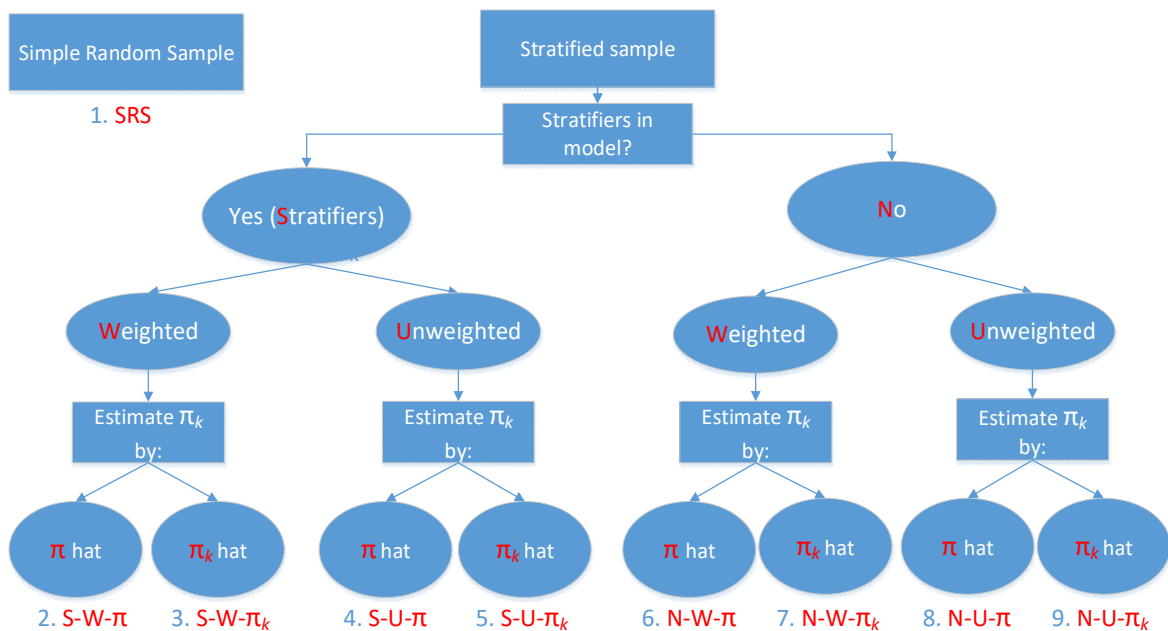


Figure 2. Nine simulation scenarios for samples drawn from the 2017 SDR PUF

Table 2 presents the number of empty and unique cells in the population and in the first replicate of the 1000 samples used in scenarios 2-5 and scenarios 6-9, respectively for the all two-way interaction models. Among the 47,520 cells overall, nearly 99% are sample zeros including about 69% population zeros and about 30% sample zeros only. The large proportion of empty cells (in particular empty population cells) is due to drawing samples from a sample (i.e., 2017 SDR PUF) in the simulation study. Among the non-zero sample cells, the majority are sample uniques. We acknowledge that the highly specific design of the

simulation study, characterized by an extremely high number of structural zeros, a large proportion of sampling zeros, and many unique cells, makes the findings difficult to generalize to real-world scenarios. We used the iterative proportional fitting (IPF) approach as used by the Skinner and Shlomo (2008) to address the structural zero issue. Any zeros in the bivariate margins (in the case of all two-way interaction models) used for the IPF are considered structural (population) zeros. In other words, any cells involved in the zero bivariate margins are considered population zeros. Since the margins are estimated using weighted sample counts, any marginal cells with population zeros are also sample zeros.

Table 2. The number of empty and unique cells in the population and in the first replicate of the 1000 samples, based on the all two-way interaction model

	Cell count	Percentage of cells
Total cells	47,520	100.0%
Population zeros	32,602	68.6%
Population uniques	2,833	6.0%
Stratified sample with stratifiers in the key variables:		
Sample zeros (including population zeros)	46,829	98.5%
Sample zero only	14,227	29.9%
Sample uniques	584	1.2%
Non-zero non-unique sample cells	107	0.2%
Stratified sample with stratifiers NOT in the key variables:		
Sample zeros (including population zeros)	46,905	98.7%
Sample zero only	14,303	30.1%
Sample uniques	474	1.0%
Non-zero non-unique sample cells	141	0.3%

4.2 Simulation Results

For simulation results, we examine the GOF against the relative difference of risk estimates RD for the 1,000 samples in each simulation scenario. The simulation results for the SRS sample (Scenario 1), the stratified sample with stratifiers in the key variables (Scenarios 2 – 5), and stratified sample without stratifiers in the key variables (Scenarios 6 – 9) are discussed in the following three subsections in

turn. We note some of the simulation runs failed to converge due to the limitation of computing resources.

4.2.1 Scenario: SRS Results

For Scenario 1 (SRS sample), the simulation results for models with main effects only and all two-way interactions are shown in Figure 3. The results show that for the independence model, all samples have both positive GOF and positive relative difference RD . The value of relative differences RD ranges roughly from 0.5 to 2.5, i.e., the estimated τ_1 (number of sample and population uniques) is about 0.5 to 2.5 times more than the true τ_1 across the 1000 samples). The average estimated τ_1 is 69 across the 1000 samples while the average true τ_1 is 33, resulting in a RD of 1.1. For the all two-way interaction model, the majority of samples have both negative GOF and negative RD , although a small proportion of the samples have positive GOF that are less than 1.96 and negative RD . Since SRS sample meets the assumptions of the log-linear models, the GOF performs as expected in general. The positive RD for the independence model and negative RD in most of the samples for the all two-way interaction models suggest that the best model lies between these two extremes. If we were to conduct the simulation to select the most accurate log-linear model according to the significant two-way interaction terms, the GOF and RD would both move closer to zero. However, the issue we are addressing in this paper is not about model search rather the variability of the weights in the complex survey design.

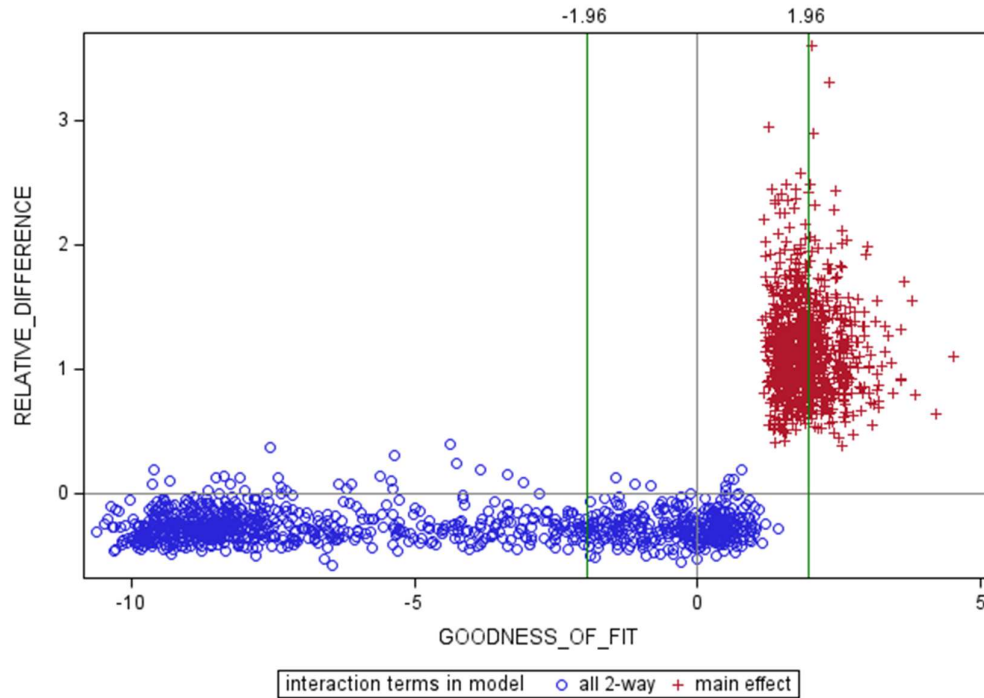


Figure 3. Relative difference (RD) of risk estimates by goodness-of-fit (GOF) statistic for 1000 simple random samples

4.2.2 Scenarios: Stratified samples with Stratifiers Included in Log-Linear Model Results

For simulation scenarios 2 – 5, we selected stratified samples and included two stratifiers in the eight key variables that will be used in the log-linear models. We looked at the independence model initially, which showed clear signs of underfitting and high disclosure risk measures (not presented here). Figure 4 shows the simulation results for the all two-way interaction models for the four simulation scenarios 2 - 5. As can be seen, for scenarios S-W- π (using weights and $\hat{\pi}$) and S-W- π_k (using weights and $\hat{\pi}_k$), almost all samples have both positive GOF and positive relative difference RD. For these two scenarios, it is evident that the models are under-fit and we will look at models with all three-way interactions next. For simulation scenario S-U- π_k (using unweighted counts and $\hat{\pi}_k$), the majority of the samples have GOF between -1.96 and 1.96 (on average -1.16 across

1000 samples) indicating good model fit and the relative difference RD mostly range from -0.2 to 0 (on average -0.1, where the average true τ_1 is 74 across the 1000 samples). For simulation scenario S-U- π (using unweighted counts and $\hat{\pi}$), over a quarter of the samples have GOF between -1.96 and 1.96 although their relative difference RD is between -0.6 and -0.8. Across the 1000 samples, the average GOF is -3.94 and average RD is -0.67 for scenario S-U- π . The magnitude of underestimation is more severe for scenario S-U- π than scenario S-U- π_k across all the samples.

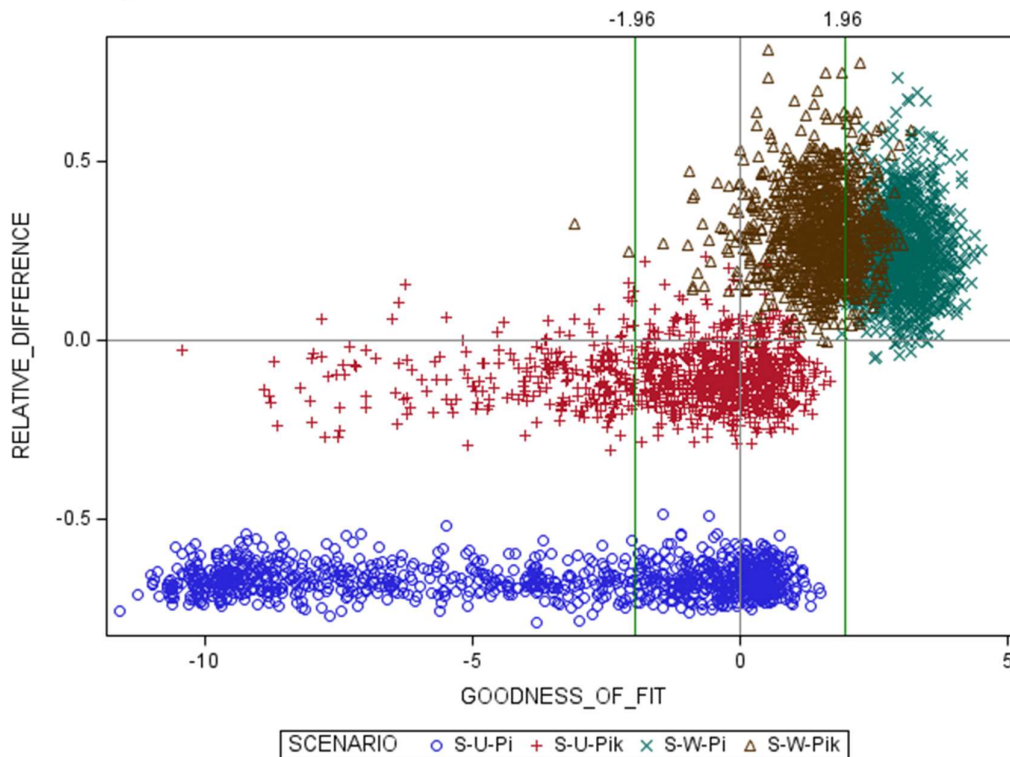


Figure 4. Relative difference (RD) of risk estimates by goodness of fit (GOF) statistic for stratified sample with stratifiers included in the log-linear model: all two-way interactions

As mentioned earlier, the models with all two-way interactions were under-fit for scenarios S-W- π and S-W- π_k . Therefore, we fit the all three-way interaction

models and show the results in Figure 5. The unweighted scenarios (S-U- π and S-U- π_k) were also included for the sake of completeness although they were not underfit with all two-way interactions. As can be seen in Figure 5, for the majority of samples in scenario S-W- π , although nearly all of the relative differences RD are negative, the GOF is greater than 1.96 indicating positive bias. This shows again that the GOF may be misleading when the model assumptions are violated. We are most concerned about the situation where the GOF is greater than -1.96 while the risk is actually underestimated. In addition, the scenario using $\hat{\pi}$ performs worse than the scenario using $\hat{\pi}_k$ (i.e., many more samples in scenario S-W- π have misleading values for GOF compared to scenario S-W- π_k). This is not surprising since $\hat{\pi}_k$ is a more accurate estimator of the cell sampling rate compared to $\hat{\pi}$ when stratifiers are used in the models. Also for the scenario S-W- π_k , nearly half of GOF is less than -1.96, indicating the all three-way interaction model is over-fitted. Satisfactory results are more likely if we were to use only the significant three-way interaction terms for this scenario. However, model-search is not the focus of this paper. Instead, the comparison between the all two-way and all three-way interactions models are used to illustrate whether the evolution from overestimation to underestimation of risk (within the variability of the criterion) is the same for all the modelling choices under the assumed scenarios.

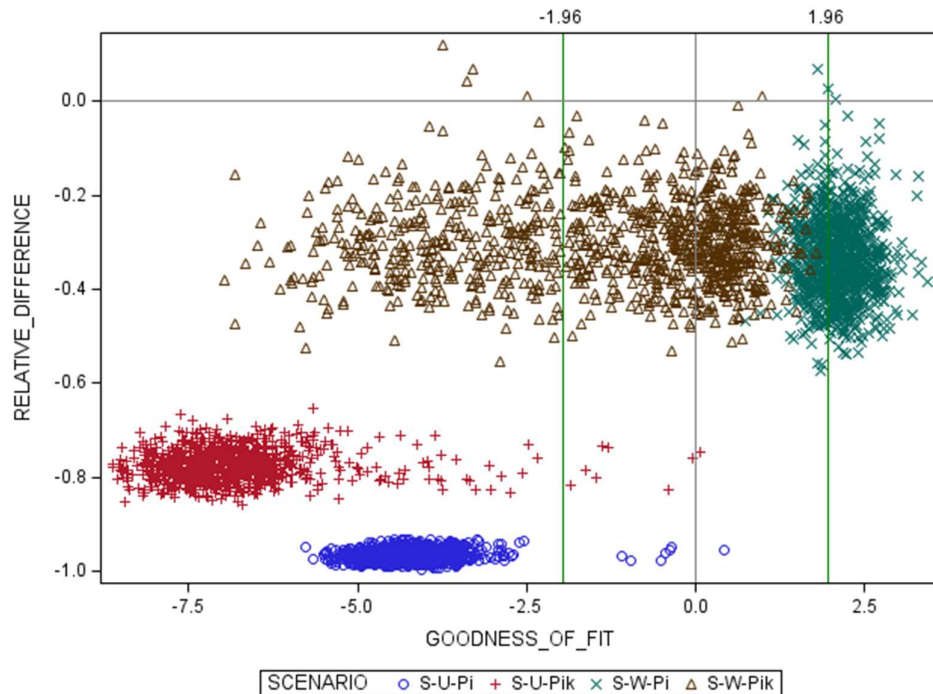


Figure 5. Relative difference of risk estimates (RD) by goodness-of-fit (GOF) statistic for stratified sample with stratifiers included in the weighted log-linear model: all three-way interactions

4.2.3 Scenarios: Stratified samples with Stratifiers Not Included in Log-Linear Model Results

For simulation scenarios 6 – 9 in Figure 2, we selected stratified samples but did not include the stratifiers in the eight key variables that were used in the log-linear models. The average true τ_1 is 27 across the 1000 samples for the four scenarios. Figures 6 shows the simulation results for log-linear models under the all two-way interactions. Compared to simulation scenarios 2 – 5 which included stratifiers in the log-linear model, the range of relative differences RD for scenarios 6 – 9 is much larger. This is because we are not able to estimate correctly the π_k and λ_k if we do not include design variables in our log-linear modelling. Among simulation scenarios 6 – 8 (N-W- π , N-W- π_k and N-U- π), the majority of samples have GOF greater than -1.96 corresponding to positive relative difference (RD) or GOF less than -1.96 corresponding to negative RD, which is as expected.

However, for scenario 8 (N-U- π_k) we see many samples have GOF less than -1.96 while RD is positive. The average GOF is -0.24 and -1.26 and average RD is 0.61 and 0.65 for scenarios N-W- π and N-W- π_k , respectively. The average GOF is -4.59 and -2.56 and average RD is -0.16 and 1.18 for scenarios N-U- π and N-U- π_k , respectively. Since it is unclear which scenario performs better, we computed the percentage of samples with GOF consistent with the relative difference RD (i.e., GOF greater than -1.96 corresponds to zero or positive difference, vice versa) as shown in Table 3. The percentage for scenarios using $\hat{\pi}_k$ (N-W- π_k and N-U- π_k) is smaller than those using $\hat{\pi}$ (scenarios N-W- π and N-U- π), which is the opposite from what we saw in Section 4.2.2 under simulation scenarios 2 – 5. Because these scenarios did not include stratifiers in the variables used in the log-linear models, $\hat{\pi}_k$ is not a reliable estimate of the cell sampling rate due to differential weights and small cell size. In addition, Table 3 shows that fitting a model with weighted sample counts perform better than fitting a model with unweighted sample counts when stratifiers are not included in the log-linear model.

Figure 7 illustrates the simulation results for log-linear models with all three-way interactions. Both unweighted scenarios (N-U- π and N-U- π_k) show clear signs of overfitting with most of the sample having GOF less than -1.96 and negative RD. For the weighted scenarios (N-W- π and N-W- π_k) it is unclear from the figure which scenario performs better. Therefore, we show the percentage of samples with GOF consistent with the relative difference RD in the rightmost column of Table 3. The percentage for the scenario N-W- π_k is higher than that for the scenario N-W- π , consistent with the results from all two-way interaction models. The percentage is similar between the two unweighted scenarios (N-U- π and N-U- π_k) because the models are so much overfitted that almost all their GOFs are less than -1.96.

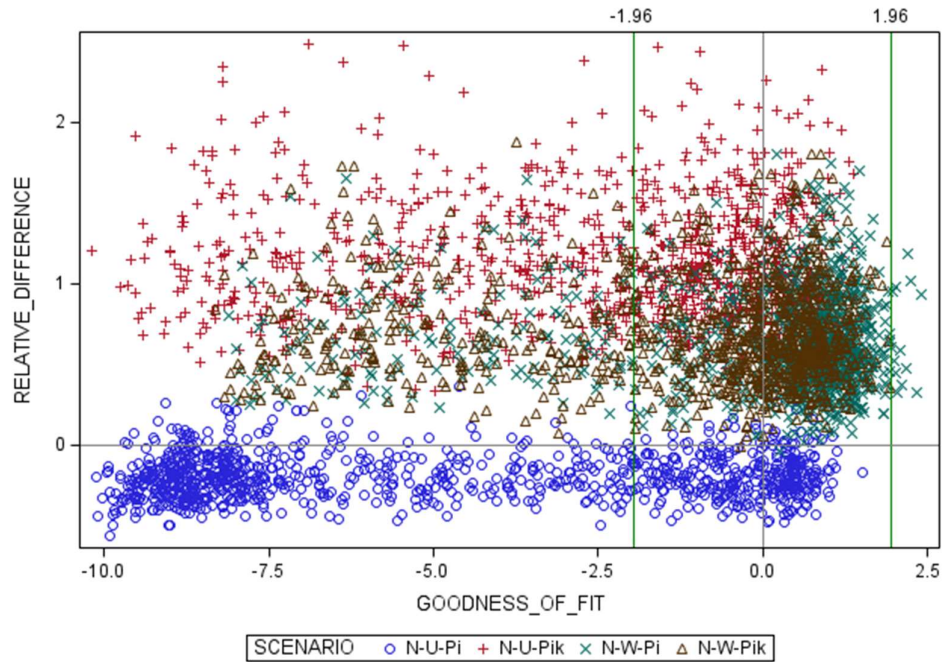


Figure 6. Relative difference of risk estimates (RD) by goodness-of-fit (GOF) statistic for stratified sample without including stratifiers in log-linear models: all two-way interactions

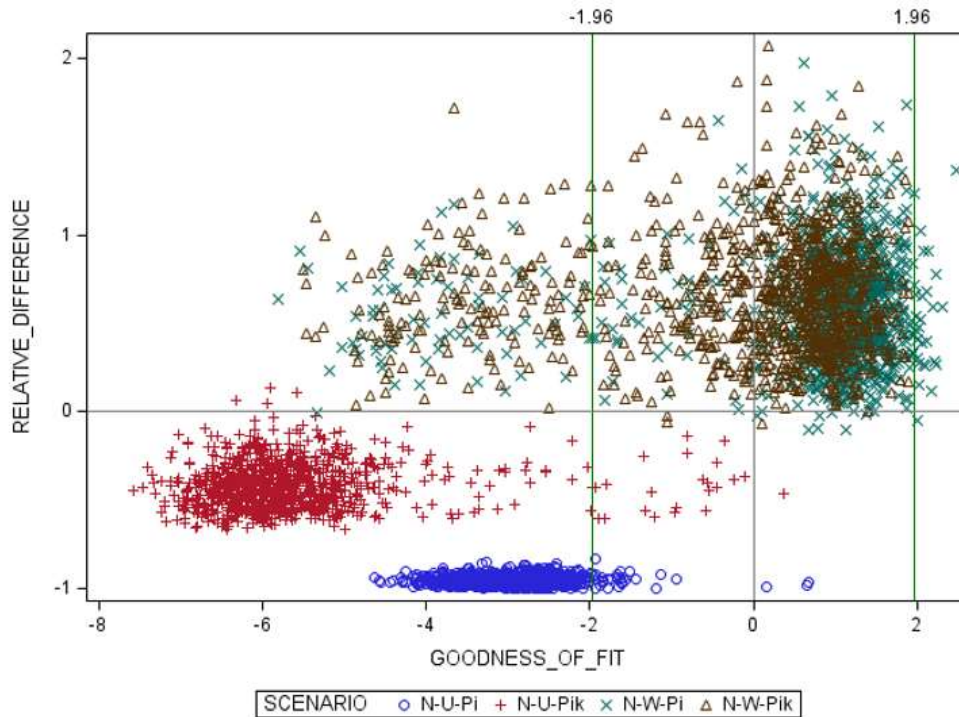


Figure 7. Relative difference of risk estimates (RD) by goodness-of-fit (GOF) statistic for stratified sample without including stratifiers in log-linear models: all three-way interactions

Table 3. Simulation scenarios 6 – 9: Percentage of samples with goodness of fit (GOF) consistent with the sign of relative difference (RD) for log-linear models under all two-way interactions and all three-way interactions

Scenario	Stratifiers in key variables	Weights used in model fitting	Estimate of sample fraction	Percent out of 1000 samples where GOF consistent with sign of RD	
				All two-way interactions	All three-way interactions
6 N-W- π	No	Yes	$\hat{\pi}$	83%	92%
7 N-W- π_k	No	Yes	$\hat{\pi}_k$	73%	84%
8 N-U- π	No	No	$\hat{\pi}$	61%	97%
9 N-U- π_k	No	No	$\hat{\pi}_k$	55%	97%

5 Application

The simulation results in Section 4 showed that GOF performed better when design variables were included in the log-linear modelling. To verify the simulation findings in a real application, we fitted log-linear models using the 2019 cohort only (i.e., the new cohort sample entering in 2019) of the 2019 SDR sample restricted use file (RUF). We included only the 2019 cohort of the 2019 SDR RUF because it was selected with a new stratification scheme and the corresponding sampling strata were available for us to use. The strata for the other cohorts in 2019 SDR RUF are not available, as was the case for 2017 SDR PUF. The sampling strata for the 2019 cohort consist of 308 levels by crossing the 77-level detailed field of study with gender (2 levels) and under-represented minority status (URM, 2 levels). Sampling weights were used to estimate the disclosure risk of re-identification. The CV of the weights for the 2019 cohort is about 78%, which is lower than the CV of weights for the 2017 SDR PUF at about 109%. Three runs of log-linear models were conducted by including different levels of complete information about the sampling strata. Table 3 shows the following for each log-linear model run: the model covariates, average cell size, GOF for the $\hat{\pi}_k$ and $\hat{\pi}$ approaches separately, and the ratio of risk estimates between these two approaches. We note that we are not able to provide the disclosure risk estimates themselves due to confidentiality reasons. The values of GOF and risk estimates are based on models with all two-way interactions and models with all three-way interactions.

From Table 4, Run 1 included gender and race/ethnicity (which can be used to derive URM), but not field of study. Run 2 replaced employer sector (EMSECDT) by the field of study with 26 categories (NSDRMENTOD). Run 3 replaced the 26-category field of study by the one with 77 categories (NSDRMEDTOD). In summary, Run 1 aggregated the 308 strata into 4 strata (gender by URM), Run 2 aggregated them into 104 strata (gender by URM by 26-category field of study), and Run 3 provided the complete 308 strata (gender by URM by 77-category field of study). By comparing results of the three runs, we can see how sensitive the GOF is to different levels of information on design variables.

Based on models with all two-way interactions, for Run 1, the GOF with the $\hat{\pi}$ approach was 3.16 indicating under-fit (and over-estimation of risk), while the GOF with the $\hat{\pi}_k$ approach was 1.28 indicating a good model fit and less bias in the disclosure risk estimates. However, the ratio of risk estimates between $\hat{\pi}$ and

$\hat{\pi}_k$ was less than 1, i.e., the risk estimate from the $\hat{\pi}$ approach was smaller than the estimate from the $\hat{\pi}_k$ approach. This pattern continued for Run 2 and Run 3. For Run 2, when the 26-category field of study was included in the all two-way interaction model, the GOF using $\hat{\pi}$ increased to almost 6 (indicating under-fit and over-estimation of disclosure risk), while the GOF for the $\hat{\pi}_k$ approach was slightly over 2 (indicating a better fit of the log-linear model). Finally, for Run 3, the GOF using $\hat{\pi}$ was still almost 6 (indicating under-fit and over estimation of disclosure risk), while the GOF for the $\hat{\pi}_k$ approach dropped to slightly under 0 (indicating a good fit of the log-linear model). Similar to Run 1, the disclosure risk estimates for Run 2 and Run 3 using the $\hat{\pi}$ approach were smaller (although their GOFs indicate over- estimation) compared to the disclosure risk estimates from the $\hat{\pi}_k$ approach.

For models with all three-way interactions, the ratio of risk estimates between $\hat{\pi}$ and $\hat{\pi}_k$ was also less than 1 for all three runs, i.e., the risk estimate from the $\hat{\pi}$ approach was smaller than the estimate from the $\hat{\pi}_k$ approach.

The poor performance of $\hat{\pi}$ approach for models with more complete design information confirms what we found in the simulation, i.e., the $\hat{\pi}_k$ approach performed better when strata were included in the model. Furthermore, with the $\hat{\pi}_k$ approach, GOF seems to perform well as long as some design information are included in the log-linear modelling, although it performed best when complete design information is included.

Table 4. Log-linear model runs for the 2019 cohort of 2019 SDR RUF

Variable name	Label	Run 1	Run 2	Run 3
ACADADMN	Academic position: dean or president, 3 categories	√	√	√
AGEGRP	Age group: 5-year intervals (bottom coded), 10 categories	√	√	√
BTHUS	Place of birth: U.S./Non-U.S., 2 categories	√	√	√
CHTOTPB	Total number of children: 1 child/2 or more children, 3 categories	√	√	√
GENDER	Gender, 2 categories	√	√	√
GOVSUP	Federal government support indicator, 3 categories	√	√	√
EMSECDT	Employer sector, 9 categories	√		
RACETHMP	Race/ethnicity, 5 categories	√	√	√
NSDRMENTOD	Field of study for first U.S. S&E or health PhD, 26 categories		√	
NSDRMEDTOD	Field of study for first U.S. S&E or health PhD, 77 categories			√
Average cell size:		0.13	0.05	0.02
Models with all two-way interactions	GOF for $\hat{\pi}$ approach:	3.16	5.87	5.64
	GOF for $\hat{\pi}_k$ approach:	1.28	2.12	-0.13
	Ratio of risk estimate using $\hat{\pi}$ approach to the risk estimate using $\hat{\pi}_k$ approach:	0.87	0.92	0.95
Models with all three-way interactions	GOF for $\hat{\pi}$ approach:	2.43	2.59	-2.99
	GOF for $\hat{\pi}_k$ approach:	0.10	-1.07	-11.61
	Ratio of risk estimate using $\hat{\pi}$ approach to the risk estimate using $\hat{\pi}_k$ approach:	0.83	0.90	0.90

6 Conclusions

We raised three research questions in the introduction in Section 1 and searched for answers through applying theory in practice and conducting simulations, as described in Section 4 with an application in Section 5. We summarize our findings to the three research questions in turn.

Research question 1: What are the likely violations of the log-linear modeling assumptions when implementing in sample surveys with complex designs?

The log-linear modeling and goodness-of-fit criteria approach developed by Skinner and Shlomo (2008) provides a scientific means to quantify the individual-level and file-level risk of re-identifying individuals in published microdata files from surveys under the disclosure risk scenario that intruders can link the survey data to other population data containing common quasi-identifying information. However, the validity of this approach relies on certain assumptions. It assumes that the sample is drawn by Bernoulli sampling and individuals in each cell formed by the model covariates have the same known inclusion probability. In applications, this assumption may be seriously violated when the survey design variables are not appropriately incorporated into the log-linear modelling (either because the design variables are not available, or because the sample design is too complicated to be accounted for). It is a logical consequence that complex survey design and unequal weighting will cause violations of the log-linear modeling assumptions. However, it does not mean that practical interventions on the survey design and weighting should be done to better meet the model assumptions.

Research question 2: For common violations of assumptions, is the GOF statistic still a good indicator of disclosure risk estimates being too high or too low? If not, what is the impact on the bias associated with the risk estimates? How can the bias be reduced?

As shown in the simulation results in Section 4, unless the sampling rates are similar across all sampled units (indicated by a small CV of the survey weights), the stratifying variables of the survey must be included as one of the model covariates in the log-linear modelling, even when the stratifiers do not seem to be highly disclosive. If a stratifier is not available, some variables related to the stratifier can be included in the model. For a survey with a large CV due to unequal selection probabilities in strata, if the stratifiers are not included in the log-linear model, the GOF can be misleading. Using the misleading GOF as guidance to select the log-linear model may result in biased disclosure risk

estimate. It would be helpful to check the robustness of the risk estimates through a sensitivity analysis. The simulation study itself illustrated a possible sensitivity analysis by using four different approaches to estimate the cell sampling rate and fit the log-linear model, and including more or fewer interaction terms into the model. Given the importance of allowing for complex sample designs and stratifying variables that may violate the assumptions of the probabilistic modelling, we recommend conducting sensitivity checks and being conservative in the disclosure risk assessment.

Research question 3: What guidance can be given to achieve a better estimate of file-level re-identification risk in survey microdata?

When the stratifiers are used as a model covariate in the log-linear modelling, using $\hat{\pi}_k$ to estimate cell sampling rate works more reliably than $\hat{\pi}$ for model selection and disclosure risk estimation. If the sampling rates across strata do not have a large variation, the $\hat{\pi}$ approach can be used for model selection estimation if the survey design variables are not available in the dataset. It needs to be distinguished the case where the stratifiers are known to the intruder and are released with the data file, and the case when they are not. In the former case, the risk may become high if the stratifiers are disclosive. Including stratifiers in the log-linear modelling might also cause sparsity problems in the contingency tables if there are a large number of strata. In the latter case, the $\hat{\pi}$ approach can be used for model selection estimation if the sampling rates do not have a large variation. But in the case that the CV is large and some or all the stratifiers are not available, the log-linear modeling approach should be used with caution since the estimated disclosure risk can be biased. It is recommended using a conservative risk estimate or seeking other methods for disclosure risk estimation.

We also offer some practical recommendations on how to select variables in order to provide meaningful disclosure risk estimates. First, since disclosure risk estimates can be sensitive to the number of variables and levels of each variable used in the log-linear modelling, a reasonable assumption is needed for the level of detailed information available to intruders. The model covariates may include quasi-identifiers that can be publicly available or obtainable in external sources, such as geographical variables; demographic variables (e.g., age, sex, race/ethnicity); and sensitive attributes (e.g., disability, income). Some categories of these quasi-identifiers may be combined if they are indistinguishable or may not convey useful information (e.g., combine the categories such as “unknown” and “others,” recode continuous age into 2-year intervals). If there are design

variables that lead to large variations in the sample selection probabilities, they should be included in the model variables as well. Since assuming that an intruder knows a large number of quasi- identifiers may overstate the disclosure risk, the number of variables can be reduced by choosing one variable to represent a group of similar variables. If, after the disclosure risk assessment, it is decided that the disclosure risk is too high and some treatment needs to be applied to a chosen variable for disclosure risk reduction, the same treatment needs to be applied to other similar variables as well. After deciding on the key variables used in the log-linear modelling, it is helpful to check the average cell size in the cross-tabulation of all the selected key variables. A very small average cell size may indicate that too many key variables or detailed categories are included in the model and the iterative proportional fitting process of the model will not converge. A general rule of thumb is to ensure that the average cell size is at least 0.01.

There are many relevant potential topics for future research. As mentioned in Section 2, this paper is limited to one of the two risk measures τ_1 and the GOF statistic $\hat{B}_1/\sqrt{V_R}$ with the robust variance estimate. It would be interesting to explore whether the other risk measure τ_2 and other versions of the GOF statistics with more exact variance estimates follow the same patterns of model selection to obtain unbiased risk estimates. In addition, our simulation did not search for significant interaction terms for the log-linear models to select the most appropriate log-linear model, rather we looked at the independence and all two-way or all three-way interaction models to assess the usefulness of the GOF which is the topic of this paper. The selection of significant terms, although computer intensive, could be explored further to find the best-fitted log-linear model.

Our simulation looked at stratified simple random samples, which may be used in school surveys, teacher surveys and business surveys. However, in household surveys cluster samples are often used instead. It would be helpful to investigate cluster sampling designs in the future. Finally, this paper examined the bias of re-identification risk estimates. An interesting topic for future research would be to estimate variance for the risk estimates as well.

Acknowledgements

This work was supported by the National Center for Science and Engineering Statistics (NCSES). The authors are grateful to Wan-Ying Chang (NCSES) for supporting this work and providing helpful comments. The authors also greatly appreciate the valuable and helpful comments from Natalie Shlomo (University of Manchester) on draft versions of the article.

References

- [1] J. Abowd, R. Ashmead, S. Garfinkel, D. Kifer, P. Leclerc, A. Machanavajjhala, B. Moran, W. Sexton, and P. Zhuravlev. "Census TopDown Algorithm: Differentially Private Data, Incremental Schemas, and Consistency with Public Knowledge," Tech. rep. US Census Bureau, 2019. <https://github.com/uscensusbureau/census2020-das-2010ddp>
- [2] Bethlehem, J., Keller, W., and Pannekoek, J. (1990), "Disclosure control of microdata," *Journal of the American Statistical Association*, 85, 38–45. <https://doi.org/10.2307/2289523>
- [3] Bishop, Y., Fienberg, S., and Holland, P. (2007), *Discrete Multivariate Analysis: Theory and Practice*, New York: Springer-Verlag. <https://doi.org/10.1007/978-0-387-72806-3>
- [4] Cameron, A. C. and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge: Cambridge University Press.
- [5] C. Carota, M. Filippone, and S. Poletti. Assessing bayesian semi-parametric log-linear models: An application to disclosure risk estimation. *Int. Stat. Rev.*, 90(1):165–183, 2022. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12471>.
- [6] Duncan, G., Elliot, M., and Salazar-González, J. J. (2011), *Statistical Confidentiality: Principles and Practice*, New York: Springer-Verlag. <https://doi.org/10.1007/978-1-4419-7802-8>
- [7] Dwork, C., McSherry, F., Nissim, K., Smith, A. (2006), "Calibrating noise to sensitivity in private data analysis," in 3rd IACR Theory of Cryptography Conference, 265–284. https://doi.org/10.1007/11681878_14

- [8] Elamir, E.A.H. and Skinner, C.J. (2006), "Record level measures of disclosure risk for survey microdata," *Journal of Official Statistics*. 22 (3), 525–539. <https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/record-level-measures-of-disclosure-risk-for-survey-microdata.pdf>
- [9] S. E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. *Ann. Stat.*, 40: 996–1023, 04 2012. doi: 10.1214/12-AOS986. URL <http://dx.doi.org/10.1214/12-AOS986>. [11] Fellegi, I. P., and Sunter, A. B. (1969), "A theory for record linkage," *Journal of the American Statistical Association*, 64, 1183–1210. <https://doi.org/10.1080/01621459.1969.10501049>
- [10] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., and de Wolf, P.-P. (2012), *Statistical Disclosure Control*, Chichester, UK: John Wiley & Sons. <https://doi.org/10.1002/9781118348239>
- [11] D. Manrique-Vallier and J. P. Reiter. Bayesian estimation of discrete multivariate latent structure models with structural zeros. *J. Comput. Graph. Stat.*, 23:1061–1079, 2014.
- [12] National Centre for Science and Engineering Statistics (NCSES) (2017) The Survey of Doctoral Recipients. Available at: <https://ncesdata.nsf.gov/datadownload/>
- [13] Poletini, S. (2003), "Some remarks on the individual risk methodology," in: *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg. <https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2003/04/confidentiality/wp.18.s.e.pdf>
- [14] Rao, J.N.K. and Thomas, D.R. (2003), "Analysis of Categorical Response Data from Complex Surveys: an Appraisal and Update," In: *Analysis of Survey Data*, eds. R.L. Chambers and C.J. Skinner, Chichester: Wiley, pp. 85-108.
- [15] Reiter, J.P. (2005) Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100, 1103-1112.
- [16] Rinott, Y. (2003), "On models for statistical disclosure risk estimation," in: *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg.
-

- [17] Shlomo, N., and Skinner, C. J. (2022), "[Measuring risk of re-identification in microdata: State-of-the art and new directions](https://academic.oup.com/jrsssa/article/185/4/1644/7069388)," *Journal of the Royal Statistical Society, Series A*, 185(4), 1644-1662.
<https://academic.oup.com/jrsssa/article/185/4/1644/7069388>
- [18] Skinner, C.J. and Holmes, D. (1998) Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14, 361-372
- [19] Skinner, C. J., and Shlomo, N. (2008), "Assessing identification risk in survey micro-data using log-linear models," *Journal of American Statistical Association*, 103 (483), 989–1001. <http://www.jstor.org/stable/27640138>
- [20] Willenborg, L. and De Waal, T. (2001), *Elements of Statistical Disclosure Control in Practice*, Lecture Notes in Statistics, 155, New York: Springer-Verlag.
<https://doi.org/10.1007/978-1-4613-0121-9>