# An Approach to Estimate the Re-identification Risk in Longitudinal Survey Microdata

Jianzhu Li[1], Lin Li[1], Tom Krenzke[1], Wan-Ying Chang[2]
[1]Westat, 1600 Research Blvd., Rockville, MD 20850
[2]National Center for Science and Engineering Statistics, National Science Foundation, 2415 Eisenhower Avenue, Alexandria, Virginia 22314

**Abstract**
Statistical agencies are committed to protecting the confidentiality of survey respondents. Prior to releasing statistical data products, a risk assessment needs to take place to ensure that the disclosure risk is at an acceptably low level. Skinner and Shlomo (2008) developed the log-linear modeling approach to measure the re-identification risk in microdata. In longitudinal surveys, because the same respondents participate in more than one wave of a longitudinal survey, the re-identification risk is usually higher than the risk in cross-sectional data. Even if the individual microdata files for each wave do not contain a common ID variable to identify the same respondents across waves, common variables that do not change over time or change in patterns may allow the users to link up the records in individual files to form longitudinal records. In this paper, we used the Survey of Doctoral Recipients (SDR) public use files as an example to demonstrate the use of log-linear modeling approach to measure the re-identification risk while incorporating the longitudinal nature of the data, which measures the increase of longitudinal risk relative to the cross-sectional risk.

**Key Words:** log-linear models, disclosure risk, risk assessment

## 1. Introduction

As an important component of survey data collection, Federal statistical agencies provide a pledge to maintain the confidentiality of respondents' data. In released data products, re-identification may occur if intruders are able to reveal the identities of respondents by linking the variables that are common in both the survey products and external data sources. As a common practice, data producers need to apply statistical and scientific principles and methods to estimate the level of re-identification risk and ensure the risk in the disseminated data is very low.

There are added complexities in assessing and limiting disclosure risk in longitudinal studies. Longitudinal surveys collect information on the same set of sampled units at different time points. A longitudinal survey may publish a single longitudinal dataset, which contains the information that is collected at each time point for the same set of respondents. A more common scenario is to release cross-sectional datasets at the end of each time point. Each cross-sectional dataset contains the respondents and their responses at a specific time point. Across time points there are respondents who participated more than once and are referred to as longitudinal respondents. If the cross-sectional datasets are linkable across different time points, more information can be obtained about the longitudinal respondents than from an individual cross-sectional dataset.

Duncan et al. (2011) summarized the data characteristics that can trigger high disclosure risk. Besides data accuracy, accessibility, comprehensiveness, level of details, and hierarchical data structures, having longitudinal records in data files is a crucial risk factor because pooling information across time may make respondents unique and therefore easily identifiable. Relative to a single cross-sectional dataset, longitudinal data contain more indirect identifying variables that can be combined to reveal respondents' identities. Some person-level characteristics that are often used as indirect identifiers (such as education, marital status, and employment status) can vary across time, so distinct patterns in these characteristics over time may increase the likelihood of identification. As for all microdata files, the existence of publicly available external data sources may trigger the disclosure risk. Especially for longitudinal surveys, linkable cross-sectional datasets may also increase disclosure risk over time as the indirect identifying information accumulates.

## 2. Re-identification Risk Estimation Method

There are different ways for quantifying and estimating the re-identification risk in microdata files. Taylor, Zhou, and Rise (2018) provide a summary of those risk measures under the statistical confidentiality protection framework. Skinner and Shlomo (2008) developed a model-based approach that can be used to quantify the re-identification risk in microdata files. The Skinner and Shlomo approach fits a log-linear model using the sample frequencies in the table cells formed by the cross-classification of a set of indirect identifying variables. The risk estimation focuses on the sample unique cases in the table cells, which are more at risk to be re-identified due to unique combinations of characteristics. The risk metrics include the expected number of sample uniques that are also population uniques, and the expected number of correct matches to population units for sample uniques. Diagnostic statistics were developed to facilitate model selection to avoid overfitting or underfitting.

We have adopted the log-linear modeling approach to estimate re-identification risk in many survey microdata files. In this paper, we demonstrate and discuss the use of log-linear models to quantify disclosure risk for survey data with longitudinal features through a case study.

## 3. Application of Risk Estimation to a Longitudinal Dataset

The Survey of Doctorate Recipients (SDR) is conducted in cycles of about 2 years by the National Center for Science and Engineering Statistics. It provides demographic and career history information about individuals who earned a research doctoral degree in a science, engineering, or health field from a U.S. academic institution. The survey samples individuals with eligible doctorates and follows them from the year of degree award to age 76. The SDR sample is refreshed every cycle with a sample of new doctoral degree earners. The 2019 SDR sample represents 1.15 million eligible doctorates in the population.

A risk assessment was conducted before releasing the 2019 SDR public use microdata file. The 2019 SDR data contain: (1) a cohort sample of new doctorate recipients since the 2017 SDR; (2) an old cohort sample of 2017 SDR respondents; (3) and a supplemental sample that was selected from the 2015 SDR sample frame to boost the sample size. As shown in

Table 1, the old cohort, or the longitudinal respondents between the 2017 and 2019 cycles, accounts for 79 percent of all 2019 SDR respondents.

**Table 1:** Overlap Between 2017 SDR and 2019 SDR

| *2019 SDR* | *2017 SDR* | *Frequency* | *Percent* |
|---|---|---|---|
| Yes | No | 17,228 | 21.30 |
| Yes | Yes | 63,654 | 78.70 |
| Total | | 80,882 | 100.00 |

The re-identification risk in the 2019 SDR public use microdata file was estimated using the log-linear models in two perspectives. First, from the cross-sectional perspective, we assumed that the risk depends on the 2019 data itself. In other words, previously published data (i.e., in 2017 or earlier) have little or no impact on the risk. Next, we reran the risk estimation from the longitudinal perspective. An analysis was conducted to evaluate the likelihood that information in an earlier cycle can be linked to the 2019 SDR data through common variables when study IDs and other direct identifiers are not released. We also revised the model specification to incorporate information from the prior cycle into the risk assessment.

### 3.1 Cross-sectional Risk Estimation

For cross-sectional risk estimation, we fitted a log-linear model with a set of indirect identifying variables from the 2019 SDR data. It should be noted that log-linear models are used to address re-identification risk only; they do not address other types of disclosure like attribute disclosure or inferential disclosure. The results of risk estimation rely heavily on assumptions—more specifically, on the amount of accurate information that is available to intruders. We typically include 6 to 10 indirect identifiers in the models, assuming that in the absence of coalition with other intruders, that is the amount of information that the intruders may know accurately. Intuitively, the more information the intruders know accurately about the respondents, the higher the risk.

Estimation of the re-identification risk focuses on sample uniques that are also population uniques, with sample uniques being defined by the set of selected key indirect identifiers. Other cases in the sample may also have re-identification risk but would be much smaller and assumed ignorable. The models do not investigate the likelihood of attack by intruders and simply assume a 100 percent chance of attack, which is not necessarily true in reality. The estimated risk from the cross-sectional analysis can be used as a benchmark for comparison with longitudinal risk.

Table 2 shows the set of indirect identifiers that were used to fit the log-linear model using the 2019 SDR. They include basic demographic variables (age group, gender, race/ethnicity, place of birth), education (year of highest degree award), family structure (number of children), and academic achievements (academic position, Federal government support). There are other indirect identifiers in the data file, but this subset was selected to represent the typical amount of information that intruders may have. With model fitting, selection and diagnostics, the re-identification risk was estimated to be about 0.7 percent, which indicated that the expected number of sample uniques that were estimated to be population uniques account for 0.7 percent of the sample size. At the file level, the average probability of re-identification is very low. The risk measure tells us the likelihood among the sample cases to be re-identified. In reality, the risk may be lower than estimated. A couple of risk-reducing considerations are: 1) data may contain errors in the survey and in external sources for various reasons; 2) there may be a mismatch in the time of the interview and the time of the data in the external source; and 3) this risk measure assumes an attack is imminent, but the probability of an attack may be far lower than one. Also, there is variance associated with the estimate, and intruders may have more or less information. This percentage should be used as a guideline for the amount of disclosure protection treatments, instead of being used for the purpose of locating population uniques.

**Table 2:** Indirect Identifying Variables Involved in Re-Identification Risk Assessment Models for 2019 SDR

| *Variable* | *Number of categories* |
|---|---|
| Academic position | dean or president, others |
| Age group | 5-year intervals (bottom coded) |
| Place of birth | U.S./Non-U.S. |
| Total number of children | no child/1 child/2 or more children |
| Gender | male or female |
| Federal government support | yes or no |
| Year of award of highest degree | 5-year intervals; bottom coded |
| Race/ethnicity | 5 categories |

### 3.2 Longitudinal Risk Estimation

Is it reasonable to assume that intruders would not be able to attach the 2017 SDR variables to the 2019 records for longitudinal respondents? Since 2017, the original respondent study ID has been replaced by a randomized ID in the SDR public use files. It is no longer possible to merge the cross-sectional datasets by common ID. However, it does not rule out the possibility of linking the longitudinal respondents in the cross-sectional datasets through other common variables, especially those that do not change values across time. To assess the likelihood of linking the longitudinal respondents through common variables, we performed a matching exercise using the 2017 and 2019 data. First, we chose the matching keys among the variables whose values remain mostly unchanged between 2017 and 2019. We identified 12 variables that have less than 200 cases that changed responses, as shown in Table 3. In addition, a 5-year age group was included in the set of matching keys, although it changed for about 40 percent of the longitudinal cases due to the natural progression of age.

**Table 3:** Matching Keys Used for the SDR 2017 and 2019 Data

| *Variables* |
|---|
| Year of award of first U.S. S&E or health PhD (5-year intervals) |
| Fine field of study for first U.S. S&E or health PhD (TOD) |
| Location of school awarding first U.S. S&E or health PhD degree (region code) |
| Citizenship/visa status at time of doctorate (recoded, from DRF), 4 levels |
| Post-graduation location at time of doctorate (from DRF): U.S., non-U.S. |
| Gender |
| Location of school awarding highest degree (region code) |
| Field of study for highest degree (minor group) (recoded for public use) |
| Year of award of highest degree (5-year intervals) (recoded for public use) |
| Place of birth (U.S./Non-U.S.) |
| Race/ethnicity (recoded for public use) |
| Location of school awarding most recent degree (region code) (recoded for public use) |

Using the matching variables, we paired up each longitudinal respondent in 2019 to all the respondents in 2017. For each pair we computed the number of differences in the matching keys between 2017 and 2019. Any non-missing values that were different between the two years were counted as differences. For each longitudinal respondent in 2019, the 2017 respondent with the smallest number of differences among the matching keys was considered as the best match. If there was more than one 2017 respondent with the same smallest number of differences, then all such cases were considered best matches. If the true match was among the best matches, the matching probability was computed as the inverse of the number of best matches. Otherwise, the matching probability was set to 0. Among the 63,654 longitudinal cases, 14,673 (23%) cases were matched with certainty (i.e., matching probability being 1). The overall number of expected correct matches (computed as the sum of matching probability) is 23,793, which is about 37 percent of the longitudinal cases.

The matching exercise shows that the risk of identifying individual longitudinal cases is not negligible even without the common case ID. It is therefore necessary to incorporate some 2017 SDR information into the risk assessment. To assess the re-identification risk due to longitudinal data, we reran the risk assessment model that was done on 2019 data, but added change indicators derived from variables that were in both the 2017 and 2019 files and that were more susceptible to changing values over the years. The variables from which the change indicators were derived are academic position, number of children, Federal government support, and year of highest degree, as shown in Table 4. For these variables, we created binary change indicators, which take the value of 1 if there were changes and 0 otherwise (the non-longitudinal respondents take the value of 0). We used these binary change indicators, instead of the 2017 variables directly, in modeling to reflect the assumed level of details and accuracy available to intruders. The variables in prior cycles do provide more information for intruders to identify longitudinal respondents, but the accessibility to prior information by intruders may become difficult as time passes. There exists a tradeoff between level of detail and accuracy in the risk evaluation.

With these change indicators being added, we reran the log-linear models on all the respondents in 2019 in two scenarios. In the first scenario, we included only the change indicators for longitudinal respondents who can be matched with certainty; in the second scenario, we applied the change indicators to all longitudinal respondents. The second scenario is more conservative and leads to a higher risk estimate. The estimated risk increased from 0.7 percent to 2.7 percent and 3.1 percent, about 4 times higher than the cross-sectional risk. The increase in risk indicates that further confidentiality protection treatments are needed. Once the treatments are applied, the re-identification risk can be re-estimated. This process can be repeated until the risk is reduced to an acceptable level.

**Table 4:** Change Indicators Involved in Longitudinal Re-Identification Risk Assessment Models for 2019 SDR

| *Variable* |
|---|
| Whether academic position changed between 2017 and 2019 |
| Whether total number of children changed between 2017 and 2019 |
| Whether Federal government support changed between 2017 and 2019 |
| Whether year of award of highest degree changed between 2017 and 2019 |

## 4. Conclusions

As shown in the SDR case study, the re-identification risk in longitudinal surveys is usually higher than the risk in cross-sectional surveys due to extra information available across time. For a longitudinal survey, risk can increase over time and becomes non-negligible when released cross-sectional datasets are linkable. If the public use files released at the end of each data collection are used for cross-sectional analysis, instead of individual-level trends, removing the common ID variable may reduce risk but not completely remove the longitudinal risk from linking data files across years. The longitudinal data contain static and dynamic variables. Static variables such as year of birth, race, gender, and place of birth allow intruders to link cross-sectional files and identify longitudinal respondents. Changes in dynamic variables such as year of highest degree, number of children, and employment status give intruders more information on the characteristics of the longitudinal respondents. For longitudinal surveys, we recommend the inclusion of change indicators for dynamic variables when using log-linear models for risk estimation, which allows us to capture an indication of the longitudinal risk caused by change over time. The use of binary change indicators in place of the previous values of the dynamic variables (e.g., use 2019 employment status and the change indicator, not the actual employment status in 2017) can partially address the loss of data detail and accuracy across time. If the information available to the intruder is overstated, the re-identification risk will be overestimated. If the data producers are interested in publishing a single longitudinal dataset to satisfy the interest of analyzing individual-level trends and trajectories, a risk assessment will need to be conducted on the longitudinal dataset, while accounting for the information that has been released earlier in cross-sectional datasets.

## References

Duncan, G., Elliot, M, and Salazar-Gonzalez, J.-J. (2011), *Statistical Confidentiality: Principles and Practice,* New York: Springer-Verlag.

Skinner, C.J., and Shlomo, N. (2008), "Assessing Identification Risk in Survey Microdata Using Log-linear Models," *Journal of American Statistical Association*, 103, 989-1001.

Taylor, L., Zhou, X.-H., and Rise, P. (2018), "A tutorial in assessing disclosure risk in microdata," *Statistics in Medicine*, 1-14.