Relative risk among respondents

In this section we investigate the relative risk among respondents with focus on records with higher risk. Figure 1 gives an overview of the general processing flow for the risk analysis that is typically conducted to investigate relative risk and the source of the risk. The process mainly includes reviewing one-way tabulations, identifying factual variables beyond the ones used in the above re-identification analysis, and processing an extensive tabulation analysis. If the risk analysis report from the extensive multi-way tabulations indicates there are no disclosure concerns in the data, no action is needed; if the report only shows some concerns, the cases with concerns can be targeted with recodes, suppression or perturbation.

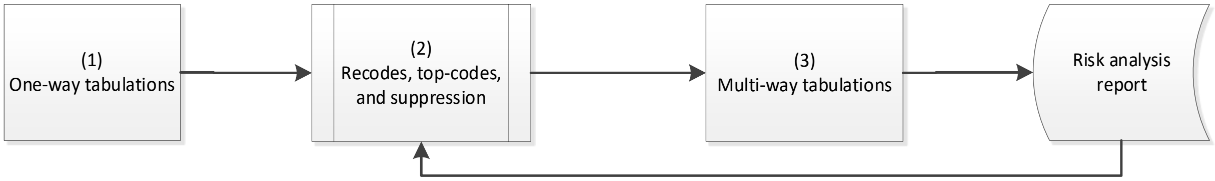

Figure 1 General processing flow for the risk analysis

One-Way Tabulations. For an initial look at the sample file, one-way tabulations of the indirect identifiers can be examined. Categories that are considered as sparse[1] at this time can be combined, and indirect identifying variables can be suppressed if a variable has two categories and one is sparse.

Multi-Way Tabulations. The primary purpose of the multi-way tabulations is to identify records with high-risk data values, and to identify causes for the high risk, such as the categories of variables that impact the risk. For example, all possible four-way tables can be processed on the indirect identifying variables. The use of four variables in a table is the number of variables that we assume an intruder would know correctly.

A table cell is defined as "with violation" when its unweighted count is less than a predetermined count. Typically a threshold rule of 3 is used to guide the risk assessment, since a count of 1 is risky

[1] For one-way tabulations, we typically consider a category as sparse if it has less than 25 respondents.

due to it being a sample unique, and a count of 2 is risky if someone in the sample can identify himself/herself in the cell, and therefore can know the characteristics of the other person in the cell. A statement can be made such as, among the 6,000 interview respondents in the sample and among all the four-way tables processed, 67 percent of respondents were involved in at least one violation.

The algorithm counts the number of violations that involve a record for the set of tables generated and groups records into multiple risk strata based on the number of violations. The proportion of table cells with violations (among all possible table cells which are of a certain dimension and involve a specific variable/category) is also computed for each category of each variable. A table can be produced to provide the percentage of cells with violations for the top 25 categories of variables with highest risk for the multi-way tables.

Combining the matching results, file risk and relative risk results leads to recommended data edits for the file. The recommendations should strive to protect respondents' confidentiality while retaining data utility as much as possible.

Note that the assessment, as with typical risk assessments, assumes all indirect identifying variables have the same "identifiability". For example, while age may be thought of as being highly identifiable, in this assessment, age is assumed to have the same identifiability as marital status.

A note about continuous variables

Data files may contain continuous variables such as age, income, and enrollment size. Top-coding is usually used to reduce the disclosure risk for continuous variables. One approach also considered is to categorize the continuous variables, since with top-coding, there is still potential for bias on computations of the average for subgroups or for regression analysis. Categorization protects against the bias and reduces disclosure risk. However, it may be considered to use top-code cutoffs for these continuous variables because the analytical value of continuous versions of some variables outweighs the potential risk. The top-coded cutoff would be used as the replacement value for the cases with values greater than the cutoff. One may also assign the average weighted value of the values above the cutpoint as the assigned value, which will lead to unbiased average for the full sample.

Accessing the Risk Quantification Software

Software for exhaustive tabulations for disclosure risk: The NCES InitialRisk SAS macro is available upon request from NCES. The R package sdcnway is available on the CRAN network at https://CRAN.R-project.org/package=SDCNway.