

## DSDR Disclosure and Confidentiality Review Policies and Procedures

This manual is as a guiding policy framework for conducting disclosure and confidentiality reviews to assess the degree of risk if or when a specific dataset is released to the public. The processes and procedures outlined are not intended to represent an exhaustive list of tasks and activities to identify disclosure and confidentiality risks, but to provide archivists, data processors, and projects managers with a set of tools and ideas on how to approach the review of datasets slated for public dissemination. Each dataset has risks unique to the specific data content and context, including, but not limited to, the original intent of the research, the type of respondent information collected, and the external data and resources available to narrow respondent identity.

What DSDR staff has developed is based on our practical experience of conducting disclosure and confidentiality reviews on several key demographic studies – reviews executed at the request of the data producers. We conducted a review of existing policies and procedures, as well as investigated how others have defined “sensitive” and/or “risky” variables or information. We also investigated and tested computer software designed to help identify disclosure risk and confidentiality issues.

### **Disclosure and Confidentiality Review Procedures**

Any review of data for disclosure risk begins with a preliminary review of the documentation and an assessment to determine which variables and information could identify a respondent. Subsequent reviews should examine Direct Identifiers; that is, information that can identify a subject or respondent directly, such as a name, address, or ZIP code.

#### Step 1 – Direct and Indirect Identifier Detection

##### *Direct Identifiers*

ICPSR reviews and documents direct identifiers present in the data that present disclosure or breach of confidentiality risk. Direct identifiers can include, but are not limited to:

- Names
- Home addresses, including ZIP Codes
- Place of work, work addresses, including ZIP Codes
- Geospatial coordinates
- Census tract number
- Latitude and longitude of home, work, etc.
- Telephone numbers, including area codes
- Social Security numbers
- Detailed income information
- Medicare/Medicaid information
- Other linkable numbers such as driver license numbers, certification numbers, etc.

The above are common direct identifiers. However, other information can be considered a direct identifier given the context of the data collection. The aforementioned list of direct identifiers is not exhaustive. Each dataset at ICPSR is reviewed for unique identifiers that could pose a disclosure risk.

##### *Indirect Identifiers*

Any disclosure review also needs to include an identification and risk assessment of variables that could provide an avenue to indirectly identify respondents. Indirect identifiers are variables that could be combined with other variables in the dataset, or with external information, to determine a respondent's identity. A common example of an indirect identification can occur in a Crosstab analysis when a cell size is very small. In particular, data are reviewed for indirect information situated within broader contextual information provided in the dataset, codebook, or other relevant documentation.

### Step 2 – Crosstab Analysis and Argus

After documentation review of direct and indirect identifiers, DSDR runs crosstab analysis to determine if particular combinations of sensitive variables can yield small enough cell size so as to identify individuals.

DSDR also uses Argus, a data analysis program that is designed to automate the process of comparing variables to detect combinations of low frequency responses that may allow an end user to identify a participant. Furthermore, Argus is also capable of automatically suppressing low frequency data that exhibits a potential disclosure risk. Argus has adjustable parameters that can be set to define an acceptable level of risk.

### Step 3 – External Linkage Investigation

DSDR then investigates the likelihood that external linking to available public datasets could lead to identity disclosure. For example, detailed race, income, and incarceration information pose an increased disclosure risk if the dataset contains information on respondents in identified cities or geographical areas.

Consider the following hypothetical situation: an unethical researcher obtains a dataset from ICPSR and decides that he would like to contact the participants with his own follow-up questions. ICPSR and/or the original researcher denies him permission to do so. The researcher decides to obtain lists of information within a given timeframe on two or more of the above-mentioned variables from publicly available databases. Using standard statistical software, he writes a simple syntax that runs a comparison between the dataset and the information he has gathered. He discovers that there is a Michael Jones who was arrested for marijuana possession on a given date, a Michael Jones who was married on a given date, and a Michael Jones who became a father on a given date. These dates correspond to the dates given on these variables by participant #123456. Therefore, it is likely that participant #123456 is Michael Jones. Even if the exact dates are not given in the data set (i.e., only month and year are given), multiple matches across several variables would indicate that the probability of a match was high. Many such matches would undoubtedly occur when large datasets are compared, and this researcher may need to locate as few as 30 participants to obtain significant results for his follow-up questions.

The main obstacle preventing this researcher from locating the participants is the difficulty in getting a list of all marriage licenses, all birth records, and all records for arrest for marijuana possession, even if it is restricted to a certain timeframe or certain city. Currently, he would have to go to each of these offices, make a written request, and pay a fee to obtain this information. These offices may not have to comply if they determine his request is not reasonable and indeed it appears that most municipal offices would not be willing to turn over such a list. However, there is no law preventing these offices from doing so – it is simply up to the discretion of each office's staff. Furthermore, it is entirely possible that at least some of these variables may be freely available on the web in as soon as the next few years. As computer technology improves, the cost of making this information available drops. Similar web-accessible databases (most notably, the sex-offender registries that most states now maintain) are already in existence. DSDR frequently reviews the type, form, and amount of such data that is available to the public to stay abreast of developments involving these possible disclosure risks.

Another potential misuse of this hypothetical dataset might come from someone the participant knows. Consider another hypothetical situation: A man with a child from a previous relationship participates in a study. The child is told about or discovers the man's participation and passes this information along to his mother. His mother suspects that the man is underreporting his income to minimize his child support payment. The mother decides to search for this study online and discovers the dataset. Using information (especially such as birth date of child, weight of child at birth, as well as date he moved, ethnic background, etc) she knows from her previous relationship with the man, she is able to determine his ID number. She can then find the income he reported during the study and check if it matches the income he reported to the family court. If it does not, she can take this information to the court and request an increase in his payments. Whether or not the court would accept this information as proof is debatable, but it still would create undue stress and inconvenience for the participant. However, it is important to keep in mind that the amount of time and effort involved in making these links decreases the probability that a participant's identity will be disclosed or confidentiality breached.

The DSDR disclosure and confidentiality review, although following a standard approach, is conducted based on the specific nature and characteristics of the data. Each dataset is reviewed for direct and indirect identifiers, and the conclusions drawn as to the disclosure risk are based on our review of the unique characteristics, information, and potential linkages of each dataset.