**De-identification and Deductive Disclosure**

Confidentiality is interested in protecting identity disclosure to unauthorized parties. The success of social science research relies on information from individuals. Thus, it is critically important to protect research participant’s identities. Disclosure risk is a term that is often used for the possibility that a data record from a study could be linked to a specific person, thereby revealing information about the person that otherwise would not be known.

Estimation of actual disclosure risk to individuals is very difficult. However, informed consent should avoid an absolute assurance of confidentiality.

---

**Key definitions**

*Confidentiality* refers to the agreements made with research participants about the information they provide will be collected, stored and analyzed. Confidentiality is comprised of two key components: direct or indirect identifying information on subjects and disclosure of information that could harm respondents.

*De-identification* is defined as the removal or replacement of personal identifiers so that it would be difficult to re-establish a link between the individual and their data. For a list of personal identifiers that should be taken out of public datasets, see page 2.

*Anonymization* is defined as the irreversible removal of the link between the individual and their data in a way that makes it impossible to re-establish the link.

One way to handle sensitive data is to restrict data access through the creation of data enclaves and data use agreement. The protected research data network is SSRI’s network space for sensitive and restricted data access.

Disclosure analysis involves the analysis of indirect identifiers that pose a rise of respondent identification. The key disclosure risk assessment component is balancing the risk and harm in disclosure with the utility of the data. O’Rourke (2011) illustrates that disclosure analyses should be part of the survey design process, and that disclosure risks should be determined early. Specifically, each variable should be analyzed to determine its statistical use, its sensitivity, its unique risk and/or how individuals might evade disclosure (e.g., issues of self-report of deviant behaviors). By doing these assessments prior to data collection, researchers could adjust the study design to mitigate some or all of these issues and also identified statistical disclosure limitation methodologies prior to use.
Key Components to be checked for Deductive Disclosure

• Identify variables that can be used to personally identify an individual. These might include personal identity components like name, address, etc. or they might include more aggregated variables such as race and gender.
• Screen all variables to determine harm risk: low / medium / high. Harm risk might include variables such as income, personal attitudes, behaviors, etc.
• Univariate screening on all variables to determine unique values.
• Targeted multivariate screening on sets of variables based on the identification of both personally identifiable data and variables that might pose a risk of harm if disclosed. This process might be iterative, based on findings.

Observations where deductive disclosure has illustrated some risk should be flagged. That might be the extent of the involvement for the data manager. However, it is also possible that the data manager might employ some of the following statistical disclosure limitation protocols.

Statistical disclosure limitation (SDL) methods focus on techniques to preserve privacy. Although there are many different techniques that have been constructed, only key ones are described here. They include: the restriction of data; data swapping or shuffling; and synthetic data.

• **Restriction of data**: Simple methods focus on restriction of the data in some way. This includes limiting the data in some way, recoding extreme values, suppression of values (typically focused on cross tabulations) and rounding data.

• **Data swapping / shuffling**: Data swapping (Moore, 1996) / shuffling (Sarathy & Muralidnar, 2002) focuses on masking the personal identities but maintain original data values through moving data randomly and attaching it to other identities.

• **Synthetic data** (Rubin, 1993; Reiter, 2005): Viewing sensitive data as missing, researchers replace the values with an appropriate posterior predictive distribution using multiple imputation techniques.
Key components within deidentification

De-identified data will not include:

- Names
- All geographic subdivisions smaller than a State. This includes:
  - Street address
  - City
  - County
  - Precinct
  - Zip code, and their equivalent geocodes
- All elements of dates (except year) for dates directly related to an individual. This includes:
  - Birth date
  - Admission date(s)
  - Discharge date(s)
  - Date of death
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code

Limited data sets can include certain identifiable data. This might include:

- Creation of age from date of birth
- Limited zip code data (first three digits):
  - The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people
  - The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000.
References


