Many national statistical agencies, survey organizations, and researchers—henceforth all called agencies—collect data that they intend to share with others. Wide dissemination of data facilitates advances in science and public policy, enables students to develop skills at data analysis, and helps ordinary citizens learn about their communities. Often, however, agencies cannot release data as collected, because doing so could reveal data subjects' identities or values of sensitive attributes. Failure to protect confidentiality can have serious consequences for agencies, since they may be violating laws or institutional rules enacted to protect confidentiality. Additionally, when confidentiality is compromised, the agencies may lose the trust of the public, so that potential respondents are less willing to give accurate answers, or even to participate, in future studies (Reiter, 2004).

At first glance, sharing safe data with others seems a straightforward task: simply strip unique identifiers like names, tax identification numbers, and exact addresses before releasing data. However, these actions alone may not suffice when quasi-identifiers, such as demographic variables, employment/education histories, or establishment sizes, remain on the file. These quasi-identifiers can be used to match units in the released data to other databases. For example, Sweeney (1997) showed that 97% of the records in a medical database for Cambridge, MA, could be identified using only birth date and 9-digit ZIP code by linking them to a publicly available voter registration list.

Agencies therefore further limit what they release, typically by altering the collected data (Willenborg and de Waal, 2001). Common strategies include those listed below. Most public use data sets released by national statistical agencies have undergone at least one of these methods of statistical disclosure limitation.

Aggregation. Aggregation reduces disclosure risks by turning atypical records—which generally are most at risk—into typical records. For example, there may be only one person with a particular combination of demographic characteristics in a city, but many people with those characteristics in a state. Releasing data for this person with geography at the city level might have a high disclosure risk, whereas releasing the data at the state level might not. Unfortunately, aggregation makes analysis at finer levels difficult and often impossible, and it creates problems of ecological inferences.

Top coding. Agencies can report sensitive values exactly only when they are above or below certain thresholds, for example reporting all incomes above \$200,000 as "\$200,000 or more." Monetary variables and ages are frequently reported with top codes, and sometimes with bottom codes as well. Top or bottom coding by definition eliminates detailed inferences about the distribution beyond the thresholds. Chopping off tails also negatively impacts estimation of whole-data quantities.

Suppression. Agencies can delete sensitive values from the released data. They might suppress entire variables or just at-risk data values. Suppression of particular data values generally creates data that are not missing at random, which are difficult to analyze properly.

Data swapping. Agencies can swap data values for selected records—for example, switch values of age, race, and sex for at-risk records with those for other records—to discourage users from matching, since matches may be based on incorrect data (Dalenius and Reiss, 1982). Swapping is used extensively by government agencies. It is generally presumed that swapping fractions are low—agencies do not reveal the rates to the public—because swapping at high levels destroys relationships involving the swapped and unswapped variables.

Adding random noise. Agencies can protect numerical data by adding some randomly selected amount to the observed values, for example a random draw from a normal distribution with mean equal to zero (Fuller, 1993). This can reduce the possibilities of accurate matching on the perturbed data and distort the values of sensitive variables. The degree of confidentiality protection depends on the nature of the noise distribution; for example, using a large variance provides greater protection. However, adding noise with large variance introduces measurement error that stretches marginal distributions and attenuates regression coefficients (Yancey *et al.*, 2002).

Synthetic data. The basic idea of synthetic data is to replace original data values at high risk of disclosure with values simulated from probability distributions (Rubin, 1993). These distributions are specified to reproduce as many of the relationships in the original data as possible. Synthetic data approaches come in two flavors: partial and full synthesis (Reiter and Raghunathan, 2007). Partially synthetic data comprise the units originally surveyed with some subset of collected values replaced with simulated values. For example, the agency might simulate sensitive or identifying variables for units in the sample with rare combinations of demographic characteristics; or, the agency might replace all data for selected sensitive variables. Fully synthetic data comprise an entirely simulated data set; the originally sampled units are not on the file. In both types, the agency generates and releases multiple versions of the data (as in multiple imputation for missing data). Synthetic data can provide valid inferences for analyses that are in accord with the synthesis models, but they may not give good results for other analyses.

Statisticians play an important role in determining agencies' data sharing strategies. First, they measure the risks of disclosures of confidential information in the data, both before and after application of data protection methods. Assessing disclosure risks is a challenging task involving modeling of data snoopers' behavior and resources; see Reiter (2005) and Elamir and Skinner (2006) for examples. Second, they advise agencies on which protection methods to apply and with what level of intensity. Generally, increasing the amount of data alteration decreases the risks of disclosures; but, it also decreases the accuracy of inferences obtained from the released data, since these methods distort relationships among the variables. Statisticians quantify the disclosure risks and data quality of competing protection methods to select ones with acceptable properties. Third, they develop new approaches to sharing confidential data. Currently, for example, there do not exist statistical approaches for safe and useful sharing of network and relational data, remote sensing data, and genomic data. As complex new data types become readily available, there will be an increased need for statisticians to develop new protection methods that facilitate data sharing.

Reprinted with permission from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science + Business Media, LLC

References

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference* **6**, 73–85.

Elamir, E. and Skinner, C. J. (2006). Record level measures of disclosure risk for survey microdata. *Journal of Official Statistics* **22**, 525–539.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9**, 383–406.

Reiter, J. P. (2004). New approaches to data dissemintation: A glimpse into the future (?). *Chance* **17**, 3, 12–16.

Reiter, J. P. (2005). Estimating identification risks in microdata. *Journal of the American Statistical Association* **100**, 1103–1113.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462–1471.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Sweeney, L. (1997). Computational disclosure control for medical microdata: the Datafly system. In *Proceedings of an International Workshop and Exposition*, 442–453.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control.* New York: Springer-Verlag.

Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002). Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 135–152. Berlin: Springer-Verlag.