Direct Identifiers

The following are direct identifiers that are to be removed from information/data to be de-identified.

1. Names
2. Geographic subdivisions smaller than a province or territory, including street address, city, region, municipality, postal code except for the initial three digits of a postal code
3. Elements of dates (except year) related to an individual
4. Telephone numbers
5. Fax numbers
6. Email addresses
7. Social Insurance numbers
8. Health Card numbers
9. Medical Record or Health Plan numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. Biometric identifiers, including fingerprints and voiceprints
15. Web universal resource locators (URLs)
16. Internet protocol (IP) address numbers
17. Full-face photographic images
18. Any other unique identifying number, characteristic, or code

This list is revised from information in the U.S. *Health Insurance Portability and Accountability Act* (HIPAA). CIHR Best Practices for Protecting Privacy in Health Research (September 2005).

Variables that Might Act as Indirect Identifiers

A dataset without direct identifiers may include indirect identifiers that in combination could lead to identification. For example,

- age
- uncommon characteristics of the individual (e.g., rare health condition, number of children)
- geographic/regional location
- named facility and/or service provider
- highly visible characteristics of the individual (e.g., ethnicity, race)

If a variable might act as an indirect identifier and compromise the confidentiality of a research participant, it can be treated in a number of ways:

- Removal – eliminating the variable from the data set
- Bracketing – combining the categories of a variable
- Top-coding – restricting the upper range of a variable
- Collapsing and/or combining variables – merging the concepts embodied in two or more variables by creating a new summary variable

- Sampling – rather than providing all of the original data, releasing a random sample of sufficient size to yield reasonable inferences
- Swapping – matching unique cases on the indirect identifier, then exchanging the values of key variables between the cases. Swapping is a service that archives may offer to limit disclosure risk
- Disturbing – adding random variation or stochastic error to the variable.

For further information see Inter-University Consortium for Political and Social Research. (2005). Guide to Social Science Data Preparation and Archiving.

Additional tips for minimizing disclosure risk:

- Use weighted data; disclosure risk is reduced when weights are used to generate output
- Avoid submitting tables with small cell sizes (i.e., cells with fewer than 5 respondents)
- Restrict cross-tabular analysis to two or three dimensions
- Be cautious when using small subgroups or small areas
- Avoid listings of cases with outliers

Statistics Canada Research Data Centres. (October, 2005). Guide for Researchers Under Agreement with Statistics Canada, http://www.statcan.ca/english/rdc/pdf/researchers_guide.pdf.

Also see,
Massell, P.B. (2003) Statistical disclosure control for tables: determining which method to use. Proceedings of Statistics Canada Symposium 2003, http://www.census.gov/srd/sdc/Massell%20StatCan%20Meth%20Symp%20english.pdf