



## ***ICE Population Health Data Repository – Data Submission Guidelines***

### **Guidelines for Preserving Confidentiality of Research Participants**

The ICE Population Health Data Repository (PHDR) places a high priority on preserving the privacy of research participants and preserving the confidentiality of respondent data. All data collections received are reviewed to ensure that confidentiality is protected and that the risk of disclosure is limited. The ICE PHDR policies and procedures in this area are governed by both professional ethics and applicable regulations. Population health researchers are bound to a code of ethics that requires that data collected for research purposes is kept confidential (see, for example, Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, Social Sciences and Humanities Research Council of Canada, *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*, 1998 [with 2000, 2002, 2005 amendments]). The rights of respondents and their continued willingness to voluntarily provide answers to scientific questions underlie this code. This applies to all participants in research, from data collectors to archivists to secondary analysts who use archived data in their research.

The ICE PHDR requires that data collections be de-identified by researchers who wish to deposit data in the repository before being sent to us. There are two kinds of variables that pose a risk to participant confidentiality – direct identifiers and indirect identifiers. In order to de-identify data direct identifiers must be removed, while indirect identifiers may be treated in a number of ways.

#### **Removing direct identifiers**

Direct identifiers are variables that point explicitly to a particular individual or unit. It is important that researchers depositing data with the PHDR remove the following direct identifiers<sup>1</sup> (where applicable):

1. Names
2. All geographic subdivisions smaller than a province or territory, including street address, city, region, municipality, postal code, and their equivalent geographical codes, except for the initial three digits (Forward Sortation Area) of a postal code, if:
  - a. The geographic unit formed by combining all postal codes with the same FSA contains more than 20,000 people.
  - b. The FSA of a postal code for all such geographic units containing 20,000 or fewer people are changed to 'ANA'.
3. All elements of dates (except year) for dates directly related to an individual, including birth date, date of death; and all ages over 89 and all elements of dates

---

<sup>1</sup> U.S. Department of Health and Human Services *Health Insurance Portability and Accountability Act* Privacy Rule.

(including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.

4. Telephone numbers
5. Fax numbers
6. Electronic mail 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. Web universal resource locators (URLs)
15. Internet protocol (IP) address numbers
16. Biometric identifiers, including fingerprints and voiceprints
17. Full-face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code

### **Dealing with indirect identifiers**

Indirect identifiers can also point to an individual or unit by focusing attention on unique cases, or in combination with other variables. Examples of indirect identifiers include: organizations to which a respondent belongs, educational institution from which the respondent graduated (and year of graduation), exact occupations held, exact dates of events, detailed income, and offices or posts held by the respondent. 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<sup>2</sup>:

- 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

The ICE PHDR does not recommend swapping or disturbing through adding noise.

### **Staff support from the PHDR**

Staff at the PHDR are able to consult researchers submitting data sets on maintaining the confidentiality of respondents. The staff will also perform a disclosure review process to further minimize the risk of disclosure and will work with investigators to resolve any remaining issues of confidentiality.

---

<sup>2</sup> Inter-University Consortium for Political and Social Research. *Guide to Social Science Data Preparation and Archiving*. 2005.

### **Additional tips for minimizing disclosure risk<sup>3</sup>**

- 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
- Let ICE staff know of any possible residual disclosure

---

<sup>3</sup> Statistics Canada Research Data Centres. *Guide for Researchers Under Agreement with Statistics Canada*. October, 2005. [http://www.statcan.ca/english/rdc/pdf/researchers\\_guide.pdf](http://www.statcan.ca/english/rdc/pdf/researchers_guide.pdf)