Release of Public Use Microdata Files for NPHS? Mission ... partially accomplished!

Yves Béland, Statistics Canada, 16H - R.H.C. Bldg., Tunney's Pasture, Ottawa, Ontario,
Canada, K1A 0T6, belayve@statcan.ca

Key words: Confidentiality, cross-sectional, longitudinal survey, disclosure control.

1. Introduction

The National Population Health Survey (NPHS) is one of Statistics Canada's major longitudinal household surveys, following a panel of approximately 17,000 people every two years for up to twenty years. The objective of this multipurpose survey, which was launched in 1994, is to measure the health of Canadians and also the determinants of health over time. General health and sociodemographic information are collected for all members in the household while detailed health information is collected for the selected longitudinal panel member. The main goal of the NPHS is not only to produce longitudinal health data for the panel members but also to provide users with reliable cross-sectional health data for all household members. It should be also mentioned that one of the additional goals of the NPHS was to allow provinces to buy-in extra sample units to improve the reliability of their infra-provincial cross-sectional estimates. For wave 2, which was held in 1996, NPHS produced two cross-sectional data files and one longitudinal file, the latter including health data for two waves. In order to have survey microdata available to the public, it is part of Statistics Canada's mandate to attempt to release Public-Use Microdata Files (PUMFs) while assessing any disclosure risk.

This paper describes the strategy adopted for releasing the PUMFs, and discusses the work done to protect the confidentiality of NPHS respondents. Due to the survey's dual nature (longitudinal and cross-sectional), and confidentiality concerns, only cross-sectional PUMFs have been released at this point. It is expected that a longitudinal PUMF would substantially increase the risks of disclosure of NPHS respondents. Although the feasibility study has not been completed, it is anticipated that a longitudinal PUMF will not be released.

2. Contemporary concerns about disclosure of information

As soon as a government statistical agency intends to release information at the microdata level, the problem of protecting the confidentiality of survey respondents arises. Even if users of such microdata files agreed on clear conditions regarding the use of these files, disclosure risks are always present when releasing PUMFs. Some users may be interested in linking these files to additional sources of information and hence increase the analytical value of the data. Other users may want to harm the reputation of the agency by disclosing confidential information. Because Statistics Canada has a high regard for protecting the confidentiality of the survey respondents in order maintain their high level of confidence, the risks of disclosure of personal information on microdata files must be assessed thoroughly.

3. Summary of confidentiality principles

Identification of a survey respondent represents the main risk of disclosure. It occurs if a correct one-to-one match is found between a data record on a PUMF and an individual in the population. Such identifications are prerequisite for disclosure. To achieve the goal of identifying a survey respondent, users can either link the microdata to other sources in order to increase the amount of information on an individual or look for a particular survey respondent on the microdata file, with or without prior information. For both approaches, an identification with certainty is possible if and only if a particular set of characteristics is unique in the population (uniqueness in the sample does not imply uniqueness in the population). Obviously, disclosure risks are strongly correlated to type and number of characteristics on the PUMFs, the number of categories per characteristic, the frequency of each category and the relationships between key characteristics. Of course, the level of geographic detail on the microdata file also plays a direct role since it restricts the survey respondents to a particular area of the country, province, region, etc.

Disclosure risks on microdata files can be controlled by various methods. It is important to understand here that the risks are only controlled and not reduced to zero. The goal of disclosure control methods is to release as much useful microdata as possible for users without providing enough useable information to identify with certainty a survey respondent. The disclosure control methods can be divided into two main categories: restriction and distortion methods. Commonly used restriction methods include removing names and addresses, reducing detail, placing minimum value on the population size, recoding some characteristics (*e.g.*, outlier treatment), subsampling, and also suppressing either or both characteristics, or critical survey respondents. Although they are used to a lesser extent, distortion methods such as random perturbation (*e.g.*, noise addition), data swapping and micro-aggregation have proven to be very useful in special circumstances.

It should be emphasized that several of these methods are often needed in order to efficiently control the risks of disclosure of microdata.

4. Releasing NPHS cross-sectional PUMFs

As stated earlier, wave 2 of the NPHS, held in 1996, collected general health and sociodemographic information for all household members while detailed health information was collected for one selected member only. NPHS produced two cross-sectional master data files and one longitudinal master file, the latter including health data for two waves. These master files are for internal use only. The NPHS general master cross-sectional file includes over 210,000 records each having 174 characteristics (variables), while the NPHS health master cross-sectional file includes more than 80,000 records with 945 variables. From these master files and as part of Statistics Canada's mandate, two cross-sectional Public Use Microdata Files (PUMFs) have already been released. (Note that three provinces bought additional sample for wave 2 collection that substantially increased the sizes of the cross-sectional files.) This section describes all the steps and the disclosure control methods involved in the successful release of the PUMFs.

All NPHS variables were divided into three broad categories: direct identifiers, indirect identifiers and sensitive variables. The direct identifiers include all the variables that clearly identify the respondent such as name, address, telephone number and health insurance number. All direct identifiers were suppressed from the PUMFs for obvious reasons.

The indirect identifiers include all the variables that cannot identify an individual without some prior knowledge of the respondent. For NPHS, the following variables were identified as being indirect identifiers:

-age group
-sex
-marital status
-language spoken
-country of birth
-immigration flag
-time since immigration
-type of household
-size of household
-number of bedrooms
- all geographic and sample design variables

-dwelling ownership
-highest level of education
-income adequacy
-household income
-activity limitation
-cause of health problem
-body mass index
-most chronic conditions

All other variables are classified as sensitive. The choice in classifying a variable as being sensitive or as an indirect identifier is debatable. It depends on the level of disclosure risk tolerated. For NPHS, the indirect identifiers can be seen as the variables that are 'discriminatingly visible' variables and the rest as sensitive. Indirect identifiers include variables that could help identify a respondent. For example, for a health survey like NPHS 'visible' variables would be activity limitation or being diagnosed with asthma, a 'non-visible' variable would be number of visits to the doctor. A variable like smoking, while quite 'visible', is 'non-discriminative' due to the fact that a fairly large proportion of the population smoke.

4.1 Disclosure control methods on indirect identifiers

The indirect identifiers are the most problematic variables for all household surveys, but excluding them from the PUMF would make the file almost useless for analysts. On the other hand, these variables are often used by users for linking PUMFs to other files. Thus a great deal of energy is spent in assessing the disclosure risks of indirect identifiers.

4.1.1 Level of geography

It was first decided to look at the level of geography that would be present on the NPHS files. Depending on the subject of the survey and the amount of information to be released, an internal rule at Statistics Canada is to place a minimum value on the population size of the regions, controlling for the sampling fraction. For NPHS, the minimum population size was fixed at 80,000 persons. This rule is arbitrary and could be relaxed or strengthened under special circumstances. For example, microdata were released for Prince Edward Island (PEI), a province of slightly more than 130,000 people. However, an urban/rural flag was also released for PEI, subdividing the province into two smaller subprovincial regions. Because the rural population is spread across PEI, it was found that releasing such information would not greatly increase the disclosure risks.

On both PUMFs, microdata were released at the urban/rural level for most provinces, and at the health-region level (without an urban/rural flag) for the provinces of Ontario (23 health regions), Manitoba (5) and Alberta (5). Microdata were also released for the cities of Montreal and Vancouver. The whole country was hence subdivided into 49 regions. It should be noted that having set a minimum value of 80,000 persons on the population size forced the removal of all sample design variables from the PUMFs. (NPHS uses an area frame divided into strata and clusters defined geographically (see Tambay and Catlin; 1995)

4.1.2 Restriction of details

The univariate unweighted counts were examined for all categories of every indirect identifier. This was performed within the lowest geographic level in each province for both files, in order to pinpoint unique and visible records in the sample. The results of this analysis were compared to the profile of the population in their respective areas (based on the Canadian Census or other sources). After evaluation of the results, two possible corrective options were examined: suppression of the problematic information at the record level or grouping of problematic categories for every record on the file.

Although some corrections were done at the record level (for example, a person speaking only French in rural PEI could have been changed to 'not stated'), grouping categories is inevitable. The problematic information often represents very small domains of interest for almost every region. In fact, there are often not enough records on the file to perform valid analyses. In these cases, categories are grouped into broader categories for every survey respondent in the sample. The categories of all NPHS indirect identifiers except sex, immigration, dwelling ownership, activity limitation and chronic conditions were grouped into broader categories. It should be mentioned that there may also have been grouping on some sensitive variables, not for confidentiality purposes but for data quality purposes; the number of records in the sample being too small.

Survey respondents may not be unique with regard to particular single variables but may become so when considering multiple variables. For example, a woman who gives birth to four children may not be unique in the population, but combining this information with age (less than 25) and household income (over $200,000) may cause a disclosure risk. Thus multivariate unweighted counts were examined for all combinations of three indirect identifiers by the lowest geographic level for both files. The grouped indirect identifiers were used for this analysis.

To facilitate the evaluation of the results, the number of times that each record appears as unique in a three-way cell of any table is computed for each geographic region; this is called the *record multiplicity of uniqueness* with regard to the indirect identifiers. The rationale behind this approach is that the higher the record multiplicity of a survey respondent, the higher the risk of being unique in the population. Of course, the choice of the right indirect identifiers is critical for a valid analysis. At this point, the objective is to look at data of each individual having a multiplicity above a certain value and to take a decision on a case-by-case basis. For NPHS, all records with a multiplicity above the 95th percentile mark in each region (*i.e.*, the records with the highest multiplicity) were examined. After evaluation, it was decided to further group the categories of some indirect identifiers even more. Language spoken, place of birth, main source of household income and highest level of education were regrouped. Problematic information of some individual survey respondents was also suppressed.

4.2 Reconstruction of families

As stated earlier, general health and sociodemographic data were collected for all members in the sampled households and included on the general file. Because the household-level variables are identical for each household member, the possibility of grouping all family members together was investigated. Records with similar household characteristics were grouped and the amount of noise within each group was measured. The *noise* is defined as the 'uncertainty rate' introduced by other records on the file. The higher the noise the smaller the chance of identifying with certainty a real and valid family. The 'noise rates' observed for all geographic levels were more than sufficient and the risks of disclosing confidential information were minimal, so no control methods were used here. Being able to reconstruct entire families with certainty would have greatly jeopardized the release of the general NPHS PUMF because of the increased amount of information available to users.

Another way of reconstructing some families or at least parts of families on the general file would have been to use the sampling weights. The very complex weighting strategy for NPHS survey respondents allows two or more persons in a same family to have identical sampling weights (see Statistics Canada, 1998 for more details on the weighting strategy) because of the nature of the weight-share and the poststratification adjustments. For this reason, sampling weights of such survey respondents were distorted by adding a random noise to them, minimising the risk of reconstructing families using the weights.

4.3 Analysis of survey weights

For each NPHS cross-sectional file, a survey weight was produced at the person level. The NPHS sample was selected from an area frame under a multistage stratified sample design. The poststratification adjustment was performed by age and sex at the province level for seven provinces and at the health region-level for the three provinces that bought additional sample units. Although the design variables were removed from the PUMFs, the potential risk of re-creating the design strata using survey weights was assessed. The strategy was to simply look at the univariate distribution of the survey weights in each design stratum in order to identify extreme values. No such distribution led to an identification of a design stratum. It should however be noted that the weighting strategy included an interprovincial migration adjustment, which reduced the problem. Because some panel members moved from one province with a large population to a smaller province between the two waves, survey weights of such persons were adjusted. (See Statistics Canada; 1998.)

4.4 Linkages of microdata files

Up to this point, disclosure control methods were applied to each cross-sectional file individually. These methods ensure that the disclosure risks of identifying survey respondents using the characteristics available on a Public Use Microdata File (PUMF) were controlled. Another dimension in assessing disclosure risks of the NPHS PUMFs is the file linkages. Because NPHS has already released cross-sectional PUMFs in wave 1, and because other surveys from Statistics Canada have also released microdata for some NPHS respondents, a file linkage risk analysis was performed prior to the release of any files.

4.4.1 Problematic pairs of PUMFs

A brief description of the pairs of cross-sectional microdata files for which the linkages risks were analysed is given next.

Waves 1 and 2 NPHS PUMFs
Because of the longitudinal nature of the NPHS, the wave 1 cross-sectional PUMFs comprise essentially the same survey respondents as in wave 2. In fact, the general files included over 35,000 persons common for the two years while the health files included over 15,000 of them. Being capable of linking either the two general (waves 1 and 2) or the two health PUMFs would have substantially increased the amount of information for users as they would have health data over two years of collection for the same survey respondents.

Wave 2 Health PUMF and wave 1 NLSCY PUMF
For various reasons such as the overlap in the questionnaire content for children, NPHS was integrated with the National Longitudinal Survey of Children and Youths (NLSCY) in wave 1. At the time of the sample selection in the first wave, the selected persons aged less than 12 years old were interviewed by the NLSCY instead of the NPHS. In wave 2, these 2,022 children were dropped from NLSCY and they rejoined NPHS longitudinal panel (see Statistics Canada, 1998 for more details). Of course, those children, with some health information, were included in the wave 1 NLSCY cross-sectional PUMF along with additional sociodemographic characteristics. In addition, some family characteristics have been released on the NLSCY PUMF as well. The ability of linking wave 2 NPHS health PUMF with the wave 1 NLSCY PUMF would again increase the amount of information.

For all these pairs of microdata files, the file linkage analysis strategy adopted was the same and is described next.

4.4.2 Assessment of file linkage risks

To some extent, the assessment of file linkage risks can be quite arbitrary. Depending on the subject of the survey and the extra information users would get by linking various files, the tolerance could vary from one area to another. At Statistics Canada, there are no fixed rules applicable to every survey. It is up to survey managers to demonstrate to the internal Microdata Release Committee that disclosure risks of the survey respondents on the files are sufficiently controlled. For a survey like NPHS or other longitudinal surveys, approval from that Committee is difficult to obtain especially due to potential file linkages between waves.

For the NPHS cross-sectional PUMFs, a five-step strategy was implemented for assessing the file linkage risks of each problematic pair of PUMFs:

Step 1 - Determine a subset of matching variables
Out of all characteristics common to both files, a subset of key matching variables was identified. For the NPHS, these variables are mainly socio-demographic characteristics such as age, sex, marital status, etc.; more than 15 characteristics were included in that list. Of course, the choice of variables is debatable but it was felt that those most likely to be used were sociodemographic variables.

Step 2 - Perform various direct matches
Within the lowest geographic levels that are common to both files and using the variables identified in step 1, several possible direct matches were performed. Note that the matches were not probabilistic link attempts. The idea was to find a worst case scenario by identifying the variables that generated the highest possible match rate between the two files.

Step 3 - Assess the file linkage risk
Among all the records that matched between the two files, three kinds of matches occurred:

- valid one-record-to-one-record matches;
- invalid one-to-one matches;
- one-to-many-records matches.

The one-to-many matches are not really increasing the disclosure risks because of the factor of uncertainty involved, but the one-to-one matches (both valid and invalid) are problematic. Note that invalid one-to-one matches are sometimes called 'spurious links'.

For the NPHS, the file linkage risk was defined according to the combination of two rates: the invalid one-to-one versus the valid one-to-one match rates. (Note that the invalid one-to-one match rate can be seen as a noise rate.) For example, two files can be linked together with a 50% one-to-one match but 75% of these matches could be wrong ones. The 75% noise rate puts sufficient uncertainty in the file linkage to ensure a fairly good level of protection. (The higher the noise the better the protection.)

Step 4 - Introduce noise in the link
Based on the magnitude of the file linkage risks observed in step 3, some additional noise was added in some matches by simply grouping the categories of some variables. These categories were grouped until the noise rates were satisfactory. Although more dramatic corrective procedures could have been used (e.g. suppression of some variables), only grouping of categories was used. In fact, less than half of the selected characteristics identified in step 1 needed any grouping.

Step 5 - Perform a multivariate analysis
In order to have a complete and thorough evaluation of the file linkage risks, all valid one-to-one matched records identified in step 3 were examined more carefully. A multivariate unweighted analysis using a subset of indirect identifiers selected from the two waves was performed within the lowest geographic common area. That analysis was identical to the one described in section 4.1.2 where key transitions over the two years were also added to the list of variables. The record multiplicity of uniqueness (among all the one-to-one matches including the invalid ones) was very low, indicating that identification of a survey respondent would be almost impossible.

The five-step strategy described above for assessing the file linkage risks between the various pairs of problematic cross-sectional PUMFs has proven to be very effective. Although all details of the observed results cannot be given here for confidentiality purposes (valid and invalid one-to-one match rates for example), the Microdata Release Committee fully accepted the results of the file linkage risk analysis along with the search for unique and visible records analysis. The wave 2 NPHS cross-sectional public use microdata files (both general and health) were released in May 1998.

5. Releasing NPHS longitudinal PUMF

As shown above, disclosure risks can be controlled quite effectively when releasing cross-sectional health microdata files from two different waves out of a longitudinal survey. Moreover, because the file linkage risks between two waves of the survey were very low, more detailed geography was added to the cross-sectional files. Obviously, the release of a longitudinal PUMF, made up of such cross-sectional files with that amount of health information and geographic detail, would substantially increase the risk of disclosure.

However it was proposed to assess the disclosure risks of a modified longitudinal PUMF where all geographic details have been removed. Such a file could be seen as a Canada-level file containing slightly over 15,000 records with health data collected over two waves. As of October 1999, the NPHS longitudinal PUMF has not yet been released for various reasons.

5.1 Issues in releasing longitudinal PUMF

Following wave 1, two cross-sectional PUMFs were released by NPHS: a general file with more than 55,000 records and a health file with more than 17,000 records. The lowest released geographical information included health regions for Ontario and British Columbia (northern interior part only), census metropolitan areas (CMA) for Vancouver and Montreal and urban/rural flag indicators for the rest of the country. For wave 2, the general cross-sectional PUMF included more than 210,000 records while the health file had more than 80,000 records. Health region indicators were released for Ontario, Manitoba and Alberta, CMAs for Vancouver and Montreal and urban/rural flag indicators for the rest of the country.

The release of the wave 2 cross-sectional PUMFs was approved mainly because it was shown that enough uncertainty existed in any link attempt between the waves 1 and 2 PUMFs. Being able to link wave 1 and 2 cross-sectional files would have resulted in the creation of a 'super' longitudinal file with highly-detailed individual information with fine level of geography. Hence this would substantially increase the risk of any disclosure. Although a modified version of the longitudinal PUMF with no geography appears to greatly diminish the disclosure risks, some major concerns are still present.

It was anticipated that even a modified longitudinal PUMF could be used as a match key for linking wave 1 and 2 cross-sectional PUMFs. This danger arises because the longitudinal file is created by simply copying the waves 1 and 2 cross-sectional general and health information of the longitudinal respondent. Moreover, any sort of link of the modified longitudinal PUMF with any of the cross-sectional PUMFs would provide users with valuable additional information. Even heavy suppression of data or grouping of categories on the longitudinal PUMF would not necessarily help. If a severely truncated file could still be used as a match key, users would be able to once again create a super longitudinal file, with all the 'unsuppressed' data and fine geographical details on the cross-sectional PUMFs.

5.2 Assessment of file linkage risks

Between the longitudinal and the cross-sectional files for each of the two waves, four pairs of files were problematic (longitudinal with the health and general PUMFs of each wave). Only the two pairs involving the wave 1 cross-sectional PUMFs were examined.

The strategy to assess the file linkage risks is somewhat different than the one described in Section 4.4.2. Because of the large amount of possible combinations of variables available for linking the files, a less 'time-consuming' approach was implemented.

The strategy adopted consisted of looking at various combinations of variables on the wave 1 NPHS cross-sectional health PUMF that would make records (survey respondents) unique. If so, such combinations would then be seen as a unique match key to link the two files because the data are exactly the same on both files.

Among all the variables on the wave 1 cross-sectional health PUMF, a combination of 94 variables that made every survey respondent unique on the file was discovered. That indicated a valid one-to-one match rate of 100% without any noise between the files. Only 66 variables were needed for the general PUMF to produce uniqueness.

These results prevented the release of the wave 2 NPHS longitudinal PUMF. Obviously, a super longitudinal file made up by linking back together the two cross-sectional PUMFs would be very easy to create. As shown above, even a modified longitudinal PUMF would be a perfect match key. Because of these conclusive wave 1 results, no file linkage risk analysis was performed using the wave 2 cross-sectional files as it was expected to yield similar results.

However, there are still slight chances to release a longitudinal PUMF. Assuming that someone was able to create a super longitudinal file, the disclosure risks of such a file could be assessed. Although the chances are small, it is still possible that such file would not disclose confidential information.

6. Conclusion

Originally, the evaluation of disclosure risks of the wave 2 files (two cross-sectional and one longitudinal) were conducted in parallel. It was thought at first that such an approach would facilitate the release of all NPHS PUMFs. Some protective actions (grouping, suppressing, etc...) could have been performed on all files at the same time. However, it was found that such a global approach was very time consuming and would have substantially delayed the release of other PUMFs. It was hence decided to concentrate the efforts on the two cross-sectional PUMFs, even though it was anticipated that the release of the longitudinal PUMF could be jeopardized. At that time, the release of the two cross-sectional PUMFs was identified to be relatively more important because of the large buy-in sample of units by three provinces and their request for having fine level of geography on the PUMFs. Because the wave 2 cross-sectional PUMFs were released, the chance of releasing the longitudinal PUMF afterwards was reduced as protective actions were no longer possible on the cross-sectional PUMFs.

7. References

Statistics Canada (1998). *NPHS 1996-97 Public Use Microdata Files.* Catalogue no. 82M0009GPE.

Tambay, J.L. and Catlin, G. (1995). Sample design of the National Population Health Survey. *Health Reports,* 7, 1, 29-38.