The Weighting Strategy of the Canadian Community Health Survey

François Brisebois, Sylvain Thivierge, Statistics Canada
François Brisebois, Statistics Canada, R.H. Coats Bldg, Ottawa (ON), Canada, K1A 0T6 (brisfra@statcan.ca)

Key words: weighting, health survey, dual-frame design

1. Introduction

Facing an increasing need for more and better health information, the Canadian Health Minister's Advisory Council on Health Infostructure, the Canadian Institute for Health Information (CIHI) and Statistics Canada conducted from 1998 to 1999, an extensive set of consultations with over 500 people including health administrators, researchers, caregivers, government officials, health advocacy groups and consumers. These consultations led to the creation of the Health Information Roadmap Initiative (CIHI; 1999a, 1999b), a series of projects, the largest of which is the recently developed Canadian Community Health Survey (CCHS).

The main objective of the CCHS is to provide reliable cross-sectional information on health status and health determinants at the national, provincial and regional levels. The strategy adopted to meet this objective was to implement a biennial cycle of data collection, which for the first year consists of a health region-level survey. Although this paper focuses only on that first year survey, more information on the biennial cycle strategy can be found in Béland, Bailie, Catlin and Singh (2000).

The collection of data for this first health region-level survey started in September 2000 and is to be completed by the end of 2001. The innovative questionnaire administered to the 130,000 respondents comprises two parts: a common content section administered to all respondents, and an optional content section that can be customized to the needs of each health region (HR).

The goal of this paper is to present the weighting strategy for the CCHS by going through the different adjustments applied in order to obtain a final set of survey weights. Section 2 explains in detail the definition of the two frames, which is necessary to better understand the early steps of the strategy. Section 3 presents the weighting strategy itself and finally, section 4 discusses the areas for future developments. Note that the paper refers only to the weighting of the sample in the ten Canadian provinces. Since the survey design used in the three territories is slightly different, some modifications to the strategy are required.

2. Survey frame

An interesting aspect of the CCHS is its use of two overlapping frames to select the sample required in the 133 health regions. An area frame is used as the primary frame, with a telephone frame serving as a secondary frame. Using two survey frames allows a better coverage of the targeted population, which is defined as all persons living in private occupied dwellings who are aged twelve or older (some standard exclusions apply). Among all household survey frames available at Statistics Canada, the already well established area frame used and maintained by the Canadian Labour Force Survey (LFS) was a logical choice to act as the primary frame for the CCHS. It has the advantage of covering the same target population, and its reliable infrastructure makes it easier for the CCHS to select, update, control and reach its sample. However, some aspects of the area frame justified the use of a second frame, including: 1- the high cost of face-to-face interviews in certain regions, 2- the inability of the area frame to provide the required sample for certain HRs, and 3- the desire for a permanent and flexible infrastructure for collecting data by telephone. A telephone frame was therefore implemented to overcome these obstacles.

2.1 Area frame

The area frame, as designed for the LFS, uses a multistage stratified cluster design. Since its definition is rather complex, only a quick overview is provided in this paper in order to introduce the terminology used in the CCHS strategy. A complete description of the LFS area frame is given in the *Methodology of the Canadian Labour Force Survey* (Statistics Canada, 1998).

First, the entire country is divided into strata formed using geographic, economic and demographic information. Then, each of the strata is divided into clusters, which are the primary sampling units. These clusters are usually defined as a group of, a fraction of, or exactly one Census Enumeration Area. The first stage of the sample process consists of the selection of these clusters within each stratum using a probability proportional to size (PPS) approach. Next, within each selected cluster, a listing of all dwellings is completed from which a systematic sample is drawn. This represents the second stage of selection. The third and final stage is the selection of

people within sampled in-scope dwellings. For the CCHS area frame, either one or two persons is selected depending on the household composition; in brief, two persons are selected from large households containing members in the 12-19 years old age group. Consult Béland et al. (2000) to obtain justifications for the approach and the exact algorithm. A total of about 115,000 of the 130,000 targeted respondents were drawn from the area frame. Finally, note that although samples of dwellings expired from the LFS (rotated-out) were available to be used as the survey frame (as is done for many surveys at Statistics Canada), CCHS selected a sample of new dwellings in order to reduce respondent burden.

2.2 Telephone frame

The telephone frame originally consisted only of a Random Digit Dialing (RDD) frame of telephone numbers. However, during collection, alarmingly low hit rates were observed in some HRs, which had adverse effects on the morale of interviewers. To overcome the situation, a list frame was introduced in problematic HRs, approximately halfway through the collection period. The list frame consists of a simple list of phone numbers, obtained from *Infobase Telephone Directories, Canadian Edition*, a commercially available CD-ROM. The disadvantages of the list frame are obvious; confidential and unlisted numbers are missing, and the list can quickly be out-dated as people move. However, it increases significantly the hit rates and consequently lightens interviewers' already heavy workloads, which is thought to result into better data quality.

The RDD frame uses the Elimination of Non-Working Banks (ENWB) method, a procedure adopted by Statistics Canada's General Social Survey (Norris and Paton, 1991). A hundreds bank (the first eight digits of a ten-digit telephone number) is considered to be non-working if it does not contain any residential telephone numbers. The frame begins as a list of all possible hundreds banks and, as non-working banks are identified from various sources, they are eliminated from the frame. The banks on the frame are then grouped to create RDD strata. Within a RDD stratum, a bank is randomly chosen and a number between 00 and 99 is generated at random to create a complete, ten-digit telephone number. This procedure is repeated until the required number of telephone numbers within the RDD stratum is reached. The details behind the grouping of numbers to form RDD strata are discussed in section 2.3.

As mentioned earlier, the list frame consists of a simple list of telephone numbers. Using a conversion file, each number listed was mapped back to a HR using the postal code present on the file. Next, using the derived HR, the sample needed for each HR was drawn from the list, using simple random sampling.

Unlike the area frame, only one person from each responding household was selected for the list and RDD frames. The primary reason for that was to reduce respondent burden.

2.3 Dealing with the HR geography

Prior to the CCHS, neither frame (area and telephone) was designed to meet the HR-level geography requirements. However, HRs can be defined in terms of Census Enumeration Areas (EA). EAs were therefore used to remap each frame to the HR geography. The remapping required intensive work and in some cases involved a manual intervention.

For the area frame, since both the HR and LFS geographies are defined in terms of EAs, the derivation of the HR was obtained by doing simple conversions. For the majority of cases, the conversion appeared to be exactly one-to-one. For the few remaining cases, where EAs matched with more than one cluster, each EA was manually assigned to one of the clusters based on population counts.

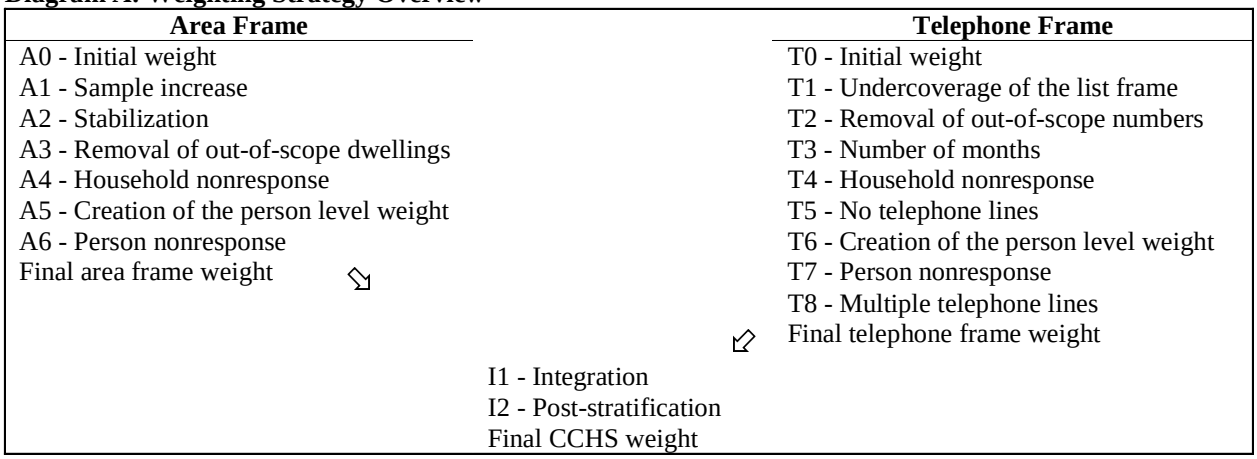
For the telephone frame, the work of assigning a HR to each sampled unit was, for both the RDD and List frame, based on the *Infobase*. Each record on the *Infobase* was assigned a HR with the help of the postal code present on the file (and using some administrative files). For the RDD frame, numbers on *Infobase* were aggregated at the Area Code Prefix (ACP) level. Within each ACP, the HR getting the majority of numbers present was the one to which the ACP was assigned. In fact, at least two-thirds of the numbers had to fall within the HR to automatically assign it as the final one; cases not meeting the two-thirds requirement were treated on an individual basis. Note that ACPs assigned to the same HR were regrouped to form a RDD stratum. For the List frame, since the sampled units were selected directly from *Infobase*, the HR derived was taken as is.

3. Weighting Strategy

The weighting strategy for the CCHS was developed by first treating both the area and the telephone frames independently to produce two sets of weights, one for each frame. These two sets of weights were then combined into a single set through a step called the integration. The following sub-sections present

each weight adjustment part of the strategy, first for the area frame, then for the telephone frame. Next follows the section on the integration and the post-stratification, which are the final steps of the weighting strategy. Diagram A presents an overview of the complete weighting strategy. It uses a numbering system that will be referred to throughout the section where each adjustment is reported in order of its application. Letters A and T are used as prefixes to refer to adjustments applied to the *Area* and *Telephone* frame units respectively, while the prefix I indicates adjustments applied to the *Integrated* set of weights.

Diagram A: Weighting Strategy Overview



3.1 Weight Adjustment - Area Frame

A0 - Initial weight
Since the mechanism established for the LFS was used to select the CCHS sample, the initial weights had to be computed with respect to that mechanism. First, clusters are selected with probabilities proportional to population sizes (based on 1991 Census counts), and then within selected clusters, dwellings are sampled using a systematic approach. The product of the probabilities for each of these selections represents the overall probability of selection, and the inverse of that probability is used as the CCHS initial weight.

A1- Sample increase
Some modifications were made to the default LFS mechanism at the time of sample selection. The current LFS design provides approximately 68,000 dwellings nationally, while CCHS requirements in terms of sample size were almost twice that number. Part of the modifications made in order to obtain the needed sample within a HR consisted of repeating the sampling process of dwellings within all clusters in the HR. This modification had the effect of boosting the sample and had to be accounted for in the weighting to correctly represent the probability of selection. The adjustment factor A1 is defined as the inverse of the number of repetitions done in the sampling process to meet the requirements.

A2 - Stabilization
In some HRs, increasing the sample as described in the previous paragraph resulted in a sample significantly larger than necessary. Stabilization was therefore instituted to bring the sample size back down to the desired level. The stabilization process consisted of randomly subsampling dwellings at the HR level, when necessary. The adjustment factor for this step is however, computed at the cluster level in order to obtain more stable adjustments, and represents the magnitude of the subsampling done.

A3 - Removal of out-of-scope dwellings
Among all dwellings sampled, a certain proportion of them is identified during collection as being out-of-scope. Dwellings demolished or in construction, vacant, seasonal or secondary dwellings, and institutions are examples of out-of-scope cases for CCHS. Records for these dwellings are simply removed from the process, leaving us with only in-scope dwellings, or equivalently referred to as households.

A4 - Household nonresponse
An adjustment is made to the weights to compensate for household-level nonresponse, that is usually when a household refuses to participate in the survey, provides unusable data, or can not be reached for an interview. Weights of household nonrespondents are distributed to respondents using response propensity

classes. The software package Knowledge Seeker (ANGOSS Software, 1995) was used to generate the classes with the help of its tree structure tool. An improved version of the CHAID (Chi-Square Automatic Interaction Detector) algorithm available in Knowledge Seeker is used to identify each node of the tree structure, based on the characteristics that best split the sample into groups that are dissimilar with respect to response/nonresponse. The final tree structure generated determines the classes to use for the weight adjustments. Since the information available for nonrespondents is limited, only characteristics such as the province, the collection period and a rural/urban indicator can be used in the creation of the classes.

A5 - Creation of the person level weight
Since the ultimate sampling unit for the CCHS is a person, the household level weights need to be converted down to the person-level using information collected from the roster of all household members. This information, combined with the number of persons selected within the household, is used to derive the adjustment factor for this step. First, each selected person is assigned the weight of his/her household. Then, depending on the number of persons selected, the number of in-scope persons aged between 12 and 19 years old, and the number of persons aged 20 years old and over, an adjustment factor is applied. For selected people from households where only one person was selected, the adjustment actually consists of the number of household members. For cases where two people aged in the 20+ age group were selected, the adjustment for each person is half the number of household members. Finally, for cases where one person in the 12-19 age group and one in the 20+ age group were selected, the adjustments are respectively the number of household members in the 12-19 age group and the number in the 20+ age group.

A6 - Person nonresponse
A CCHS interview can be seen as a two-part process. The first part consists of the interviewer getting the complete roster of the people living within the responding household, and the second is selecting the CCHS respondent to conduct the interview. In some cases, interviewers can only get through the first part, either because they can not get in touch with the selected person, or because that selected person refuses to be interviewed. Such cases are defined as person nonresponses and an adjustment factor is applied to the weights of respondents to overcome this nonresponse. Since basic socio-demographic characteristics such as the province of residence, age, sex, education, and the marital status are available for all selected people (obtained with the roster of all household members), they were used to define appropriate adjustment classes. Similarly to the household nonresponse adjustment (A4), Knowledge Seeker was used to generate the classes.

Final area frame weight
Once all adjustments are computed and applied successively, a final area frame weight is obtained. Since not all HRs are covered by the area frame (only 5 HRs are actually not covered), this weight can not be considered as a national or provincial level representative weight, and therefore needs to be combined with the telephone frame weight before doing any estimation.

3.2 Weight Adjustment - Telephone Frame

T0 - Initial weight
The initial weight is computed slightly differently between the RDD and List frame samples. Both are defined as the inverse probability of selection, but the methods of selection, and therefore the probabilities, differ. For the RDD, the selection of numbers is done within each RDD stratum. An RDD stratum is an aggregation of ACPs, each containing valid banks of one hundred numbers. Therefore, the inverse probability of selection is the ratio of the number of sampled units to one hundred times the number of banks within the RDD stratum. For the list frame, telephone numbers are selected among all numbers available on the list, within the HR for which the unit is selected. Hence, the inverse probability of selection corresponds to the ratio of the number of sampled units to the number of telephone numbers in the list within the HR.

T1 - Undercoverage of the list frame
As mentioned earlier in the paper, the list frame has the disadvantage of not covering some phone numbers, which are actually covered by the RDD frame. In order for them to be processed together through the rest of the weighting, one adjustment was necessary: adjusting the list frame weights to account for the undercoverage of that frame relative to the RDD one. The adjustment consists of inflating the weights of the list frame units by the amount of undercoverage, individually for each HR. Estimating the undercoverage was one of the most challenging part and was done using the data collected from the CCHS area frame sample. For all people interviewed via the area frame, the questionnaire includes a set of questions verifying if the household has a telephone, and how many lines it has. Phone numbers are collected for positive answers, and then matched to the Infobase to see if they are listed. The proportion

of unlisted numbers represents the desired undercoverage rates.

T2 - Removal of out-of-scope numbers
Telephone numbers leading to businesses, institutions or other dwellings out of the scope of the survey, as well as numbers not in service or any other non-working numbers, are all examples of out-of-scope cases for the telephone frame. As for the area frame, these cases are simply removed from the process, leaving only in-scope dwellings.

T3 - Number of months
Contrary to the area frame where the entire sample was selected at the beginning of the sampling process, samples were drawn monthly for the telephone frame. Each of these monthly samples comes with an initial weight that makes each sample representative at the national level. Since the process combines several monthly samples, an adjustment must be applied so that the total sample weights sum up to only one times the Canadian population.

T4 - Household nonresponse
A similar process to the one done for the area frame (A4) is applied here to adjust for household nonresponses. The same characteristics are available to help create the adjustment classes.

T5 - No telephone lines
It is believed that about 1 to 2% of the Canadian population does not have a telephone line. As explained in step T1, information about the presence of a telephone is collected for the area frame sample, which can be used to estimate the proportion of households without a phone at the HR level. Similarly as for T1, the telephone frame sample weights are inflated to account for that uncovered population based on proportions observed with the area frame data. This adjustment is applied at the HR level.

T6 - Creation of the person level weight
This adjustment converts the household level weight to a person level weight. Since only one person is selected per household for the telephone frame, the adjustment factor is simply the total number of in-scope household members.

T7 - Person nonresponse
This step is exactly the same as the person nonresponse adjustment for the area frame (A6).

T8 - Multiple telephone lines
The fact that some households can possess more than one telephone line has an impact on the weighting; having more lines translates into having a higher probability of being selected. Therefore, the weights need to be adjusted for the number of non-business telephone lines the household has. This information is obtained during the early stage of the interview.

Final telephone frame weight
Once all adjustments are applied, the remaining records are the telephone frame respondents. As for the area frame, this set of weights can however not be considered as nationally or provincially representative since not all HRs are covered by the telephone frame.

3.3 Integration and post-stratification

I1 - Integration
This step consists in integrating both sets of weights to create one single CCHS weight. The literature proposes various approaches to integrate sets of weights for dual-frame surveys; see Skinner and Rao (1996) for an overview of existing methods and other references on the topic. These approaches are generally based on effective sample sizes, that is, the ratio between the sample size and the design effect. Studies are being conducted to examine which approach should be applied to the CCHS. The main difficulty resides in the fact that external sources of HR level design effects are almost nonexistent. The best source of health survey oriented design effects is actually the CCHS itself. The studies will examine the possibility of using preliminary files (containing 6 or 9 months of collected data) to estimate the needed design effects, and their reliability at the HR level. Final decisions on the approach to use will be made at the time of production.

I2 - Post-stratification
Finally, a post-stratification is done to ensure that the final weights sum to the 2000-01 population totals at the HR, age, and sex levels. For each HR, population totals estimated using 1996 Census counts and some demographic growth information are computed for ten age-sex groups: five age groups (12-19, 20-29, 30-44, 45-64, 65+) for both males and females.

Final CCHS weight
After having applied all the necessary adjustment factors, the resulting set of weights consists of the final CCHS weights that can now be disseminated with the data files, and be used for estimation.

4. Future Developments

Another important issue involved in the weighting is the creation of share and share-link files to meet provincial governments and Health Canada needs, i.e., files containing only those respondents who agreed that their data be shared or shared and linked with other health-related files. Thus, these subsets of respondents have to be reweighted and post-stratified to population totals. Non-share and non-link classes will be created to adjust the weights and remove any potential bias caused by this type of non-response.

In addition, the weighting process will include several verifications of the weights produced (e.g., calculation of coverage rates, outlier detection) to evaluate the quality of the sampling frames, the sampling design and the estimation strategy adopted.

Finally, the bootstrap method will be used to produce variance estimates. Studies comparing several variance estimation techniques have led to the adoption of the bootstrap method in the National Population Health Survey (NPHS), a longitudinal health survey launched in 1994 by Statistics Canada. See Yeo, Mantel and Liu (1999) for more details on the bootstrap method in the context of the NPHS. The advantages of the bootstrap over other variance estimation techniques described in this paper and the similarity between the CCHS and NPHS sampling frames justified this choice for the CCHS.

This document was written while only preliminary weighting for internal use was done. The final weighting is planned for early 2002. Some modifications in the weighting strategy could be made. Readers are invited to consult CCHS documentation for a more updated version of the weighting method, once data are released in the spring of 2002.

5. Acknowledgement

The authors would like to thank their colleagues in methodology who participated in the development of the weighting strategy. They are also grateful to Vincent Dale, Jack Gambino and Marianna Morano for their insightful comments during the writing of this paper.

6. References

ANGOSS Software (1995). Knowledge Seeker IV for Windows - User's Guide. ANGOSS Software International Limited.

Béland, Y., Bailie, L., Catlin, G. and Singh, M.P. (2000). *CCHS and NPHS—An Improved Health Survey Program at Statistics Canada.* Proceedings of the Survey Research Methods Section. American Statistical Association. To be published.

Canadian Institute for Health Information (1999a), *Health Information Roadmap: Beginning the Journey*. (1-895581-32-X).

Canadian Institute for Health Information (1999b), *Health Information Roadmap: Responding to Needs*. (1-895581-30-3).

Morano M., Lessard, S. and Béland, Y. (2000). *Creation of a dual-frame design for the Canadian Community Health Survey*, 2000 Proceedings of the Survey Methods Section, Statistical Society of Canada.

Norris, D.A. and Paton, D.G. (1991). *Canada's General Social Survey: Five Years of Experience*, *Survey Methodology*, 17, 227-240.

Skinner, C.J. and Rao, J.N.K. (1996), *Estimation in Dual Frame Surveys With Complex Designs*, Journal of the American Statistical Association, 91, pp. 349-356.

Statistics Canada (1998). *Methodology of the Canadian Labour Force Survey.* Statistics Canada. Cat. No. 71-526-XPB.

Yeo, D., Mantel, H. and Liu, T.P. (1999). *Bootstrap Variance Estimation for the National Population Health Survey*, 1999 Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 778-783.