

CREATION OF A DUAL-FRAME DESIGN FOR THE CANADIAN COMMUNITY HEALTH SURVEY

Marianna Morano, Suzanne Lessard and Yves Béland¹

ABSTRACT

The choice of a sampling frame depends on many factors but, first and foremost, the frame must correspond as closely as possible to the target population of the survey. Moreover, the creation, utilization, updating and verification of the sampling frame must fall within the operational and budget constraints of the survey. For the regional component of the Canadian Community Health Survey, it was decided to use two overlapping sampling frames, the area frame established for the Canadian Labour Force Survey and a Random Digit Dialling frame of telephone numbers.

KEY WORDS : Regional survey; cross-sectional; health; area frame; Random Digit Dialling.

RÉSUMÉ

Le choix d'une base de sondage pour tirer un échantillon dépend de plusieurs facteurs mais la base doit d'abord et avant tout correspondre le plus possible à la population cible de l'enquête. De plus, la création, l'utilisation, la mise à jour et la vérification de la base de sondage doivent respecter les contraintes opérationnelles et budgétaires de l'enquête. Pour la composante régionale de l'Enquête sur la santé dans les collectivités canadiennes, il a été décidé d'utiliser deux bases de sondage chevauchantes: la base aréolaire mise en place pour l'Enquête sur la Population Active du Canada et la base de composition aléatoire de numéros de téléphone.

MOTS CLÉS : Enquête régionale; transversal; santé; base aréolaire; composition aléatoire.

1. INTRODUCTION

In order to address priority health data gaps, Statistics Canada has launched a new survey: the Canadian Community Health Survey (CCHS). The main objective of the CCHS is to provide reliable cross-sectional information in order to address data needs at the national, provincial and regional levels. In order to meet the stated needs of users, a biennial cycle of data collection was implemented for the CCHS. It consists of two survey components: a health region-level survey the first year with a total sample of more than 130,000 respondents called the *regional component*, and a province-level survey with a sample of 30,000 respondents in the second year referred to as the *provincial component* (Béland, Bailie, Catlin and Singh, 2000).

The primary objective of the regional component, which started collection in September 2000, is to provide reliable cross-sectional estimates with respect to health determinants, health status and utilization of the health system for 136 health regions (including one health region for each territory). For various reasons, but mainly to provide the required sample for every health region, two overlapping frames were used to select the sample for this regional component: an area frame used as the primary frame, and a Random Digit Dialling (RDD) frame of telephone numbers as a secondary frame. For simplicity, this paper focuses on the creation of the dual frame for the regional component of the CCHS in the ten provinces. (The CCHS uses a different design in the three northern territories.)

¹ Marianna Morano, 16-K, R.H. Coats, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, marianna.morano@statcan.ca
Suzanne Lessard, 16-J, R.H. Coats, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, suzanne.lessard@statcan.ca
Yves Béland, 16-H, R.H. Coats, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, yves.beland@statcan.ca

Section 2 gives an overview of the sample design of the regional component of the CCHS. Section 3 describes the adaptations made to the existing area frame in order to suit CCHS needs while section 4 gives details about the stratification of the RDD frame into health regions. The integration of the two frames is discussed in section 5. Finally, areas for future development are discussed in section 6.

2. SAMPLE DESIGN

2.1 Creation of the health regions

Health regions (HR) are geographical regions that the provinces use for administrative purposes in the field of health. Statistics Canada, in agreement with the provinces, has revised the boundaries of some HRs according to the geography of the 1996 Census, to allow for the production of demographic projections for different age/sex groups. For statistical purposes, 133 HRs distributed in the ten provinces are considered for the regional component.

2.2 Target population

The CCHS targets persons living in private occupied dwellings who are aged twelve or older, and covers approximately 97% of this population. Persons living on Indian Reserves or on Crown lands, residents of institutions, full-time members of the Canadian Armed Forces and residents of certain remote regions are excluded from this survey.

2.3 Sample size and allocation

To provide reliable estimates for the 133 HRs, a sample of 130,750 is desired. Because HRs vary greatly in size and number from one province to another, it is difficult to establish an equilibrium between regional and provincial needs. The strategy that has been adopted consists of three steps, giving relatively equal importance to the HRs and to the provinces (Béland *et al.*, 2000). The first two steps allocate the total sample to the provinces as a function of the number of HRs they contain and of their respective populations. At the third step, the provincial sample is allocated among the HRs proportionally to the square root of the size of the population in the regions. This three-step approach guarantees each HR sufficient sample with minimal disturbance to the provincial allocation of sample sizes. Table 1 summarizes the distribution of HRs by population size as well as the average sample size planned by HR category.

Table 1. Mean sample sizes by category of HRs

Category of HRs	Population Size	# of HRs	Mean Samp. Size
Small	Less than 75,000	41	530
Medium	75,000 – 240,000	60	900
Large	240,000 – 640,000	25	1,500
Very Large	More than 640,000	7	2,500

2.4 Sampling frame

The choice of a sampling frame depends on many factors but, first and foremost, the frame must correspond as much as possible to the target population of the survey. Another key consideration is the cost of creating and maintaining (i.e. updating) the frame. For this survey component, it was decided to use two overlapping sampling frames, the area frame established for the Canadian Labour Force Survey (LFS) and a Random Digit Dialling (RDD) frame of telephone numbers, with sample being drawn primarily from the area frame. Face-to-face interviews would be conducted for persons selected from the area frame, while telephone interviews would be conducted for those selected from the RDD frame.

Apart from the fact that the target population is the same as that of the LFS, the advantages of using the area frame set-up for the LFS for selecting the sample are undeniable. The infrastructure, which is already in place for updating new buildings, demolished buildings and excluded units, as well as the entire evaluation process of the frame coverage, are definite assets. Moreover, given that several other Statistics Canada household surveys also use this area frame, sample overlap between surveys is easier to control.

The limitations to using the RDD frame are evident: under-representation of households without telephones (~2%) or with cellular phones only (estimated to be between 1% and 2%), the generally lower response rate, and the need to make several calls before contacting a valid household. Despite these limitations, a dual frame approach is necessary for the CCHS for the following reasons: i) the high cost of face-to-face data collection in certain areas; ii) the inability of the area frame to provide the required sample for certain HRs; and iii) the desire for a permanent and flexible infrastructure for collecting data by telephone.

2.5 Sampling strategy

The majority of the sample (115,000 respondents from the targeted 130,750) is drawn from the area frame while the remaining respondents are drawn from the RDD frame.

The 115,000 respondents from the area frame come from a sample of 97,000 households. In 79,000 households, one person per household is selected; two people per household are selected from the other 18,000. A second person is selected in some households in order to increase the representativity of certain specific subpopulations (Béland *et al.*, 2000). The sample is completed by drawing one person from each of the 15,750 households from the RDD frame.

2.6 Questionnaire content

One key goal of this survey is to provide health data on issues unique to HRs. In order to achieve this goal within the maximum 45-minute interview length planned, a strategy that is innovative and unique to the CCHS was implemented (Statistics Canada, 1999).

The final questionnaire was divided into two parts – a common content section of 35 minutes in length and an optional content section of 10 minutes, customized to the HR needs. The provinces and HRs were provided with 28 questionnaire modules from which to choose, and the most popular choices determined the common content. Once the common content was decided, the provinces and HRs were asked to determine their HR optional content by choosing from among the remaining modules. This process has resulted in 27 different versions of the questionnaire. For a copy of all questions on the final questionnaires the reader is referred to http://www.statcan.ca/health_surveys.

3. SAMPLING FROM THE AREA FRAME

3.1 LFS current sampling plan

The area frame, as designed for the LFS, covers almost the entire country, from which a sample of dwellings is selected under a multistage stratified cluster design (Statistics Canada, 1998). For the purpose of the plan, each province is divided into three types of regions: major urban centres, cities and rural regions. Geographic or socio-economic strata are created within each major urban centre. Within the strata, between 150 and 250 dwellings are grouped to create clusters. Some urban centres have separate strata for apartments or for census enumeration areas (EA) in

which the average household income is high. In each stratum, six clusters or residential buildings (sometimes twelve or eighteen apartments) are chosen by a random sampling method with probability proportional to size (PPS), the size of which corresponds to the number of households. The design of the LFS allows for renewing one-sixth of the sample each month.

The other cities and rural regions in each province are stratified first geographically, then according to socio-economic characteristics. In the majority of strata, six clusters (usually census EAs) are selected using the PPS method. Where there is low population density, a three-step plan is used whereby two or three primary sampling units (PSU), which normally correspond to groups of EAs, are selected and each PSU is divided into clusters, six of which are sampled. The selection is made at each step using the PPS method.

Once the new clusters are listed, the sample is obtained using a systematic sampling of dwellings. The *yield* is the number of households selected within the framework of the LFS for a given month. As the sampling rates are determined in advance, there is frequently a difference between the expected sample size and the numbers that are obtained. For example, the yield of the sample is sometimes excessive. This happens particularly in areas where there is an increase in the number of dwellings, due to new construction, for example. To reduce the cost of collection, excessive output is corrected by eliminating, from the beginning, some of the units selected and by modifying the design weight. Such an operation, usually conducted at an aggregate level, is called *sample stabilisation*. In addition, the required sample of households is increased to account for out-of-scope dwellings (experience shows that 12% of all dwellings are not occupied by in-scope households, for example vacant or seasonal dwellings).

3.2 Adapting the LFS strategy for the CCHS

Not only is the CCHS sample chosen from the area frame established for the LFS, the mechanism for selecting LFS households is also used. Changes have been made to this selection process to meet the requirements of the CCHS in terms of sample size at the HR level.

To get a base sample of 97,000 households, 123,000 dwellings must be selected from the area frame (to account for out-of-scope dwellings and nonresponding households). The LFS design provides approximately 68,000 dwellings distributed across economic regions. The CCHS requires a total of 123,000 dwellings

distributed in the HRs, which have different geographic boundaries from those of the LFS economic regions. Overall, the CCHS requires almost twice as many dwellings than those generated by the LFS selection mechanism, or a *boost factor* of 1.8 (123/68). At the HR level, however, the boost factor varies from 0.6 to 6.0, which requires certain adjustments.

The changes made to the selection mechanism in a HR vary depending on the size of the boost factors. For HRs that have a factor smaller than or equal to 1, a simple stabilisation, as described above, is applied to the sample of dwellings. For those with a factor greater than 1 but smaller than or equal to 2, the sampling process of dwellings within a PSU is repeated for all selected PSUs that are part of the same HR. For HRs with a factor greater than 2 but smaller than or equal to 4, the PSU sampling process, as well as that of dwellings in a PSU, is repeated. For HRs with a factor between 4 and 6, the PSU sampling process is repeated not once but twice while that of dwellings is repeated only once. Where the chosen approach creates an unnecessary surplus of dwellings, stabilisation is performed.

It should be noted that the changes made to the LFS mechanism result in, at most, tripling the number of PSUs selected and, at most, doubling the number of dwellings selected in the PSUs, which explains the maximum boost factor of 6.0. At the HR level, boost factors were purposely capped at 6.0 for two reasons: to limit the listing of clusters (each new selected PSU requires a listing), and to avoid possible cluster effects created by too great a number of dwellings selected in a single PSU. This limit to the boost factor of certain HRs has consequently dictated the number of households required from the RDD frame.

4. SAMPLING FROM THE RDD FRAME

4.1 Elimination of Non-Working Banks

The sampling of households from the RDD frame uses the Elimination of Non-Working Banks (ENWB) method, a procedure adopted by the 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, they are eliminated from the frame. It should be noted that these banks are eliminated only when there is evidence from various sources that they are non-working. When there is no

information about a bank it is left on the frame (Paton and Dolson, 1998).

The banks on the frame are 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. Frequently, the number generated is not in service or out of scope, and therefore many additional numbers must be generated to reach the targeted sample size. This success rate is referred to as the *hit rate* and varies from region to region.

4.2 Stratification of the RDD frame

The existing RDD frame, as designed for the General Social Survey is stratified into 21 RDD strata, covering the ten provinces. With the exception of PEI, each province consists of a Census Metropolitan Area (CMA) stratum and a non-CMA stratum. Montreal and Toronto make up the two remaining strata.

CCHS needed a finer level of stratification. Since RDD sample would not be required in every HR and because questionnaire content could vary among HRs, it was important that the banks in a given HR be identified. The challenge in using the RDD frame for CCHS purposes lay in regrouping the working banks to create RDD strata that corresponded as closely as possible to the geographic boundaries of the HRs, to allow for the efficient control of sample selection.

After exploring various possibilities, the route chosen was to attempt to derive a level of geography from telephone numbers using various administrative files available. The Canada Phone directory, a commercially available CD-ROM consisting of names, addresses and telephone numbers from telephone directories in Canada, was linked to internal conversion files. This resulted in a HR being assigned to every published telephone number in the country. This data was aggregated at the Area Code Prefix (ACP) level, which is defined as the first six digits of a ten-digit telephone number.

If more than two-thirds of an ACP's telephone numbers map to a given HR, then that HR is said to be its *primary HR*. Design strata were formed by grouping together all ACPs that shared the same primary HR, regardless of the other HRs to which they mapped. This stratification resulted in 133 strata of this type (one for each HR). ACPs without a primary HR were dealt with on a case-by-case basis. In some cases, the two-thirds rule was relaxed, and a primary

HR was assigned. In other cases, the ACP was grouped with similar ACPs (ACPs without a primary HR mapping mostly to the same combination of HRs) to form a new "straddling" stratum. Ten strata of this kind were formed for a total of 143 RDD strata.

The coverage of HRs by stratum varies between 92% and 100%. Moreover, the overcoverage resulting from the fact that up to one third of each ACP may fall outside its primary HR is generally low. All of the 133 main strata except two have more than 90% of their numbers falling in their associated HRs (the exceptions are one HR in BC and one in PEI, with 82% and 85% respectively).

The number of telephone numbers needed by RDD stratum depends on the desired number of RDD respondents in each HR, the nonresponse rate (a nonresponse rate of 15% was assumed for RDD), and the hit rate in each RDD stratum of interest. The hit rates range from 15% to 61% in the RDD strata in which sample is required. In total, a sample of over 51,000 telephone numbers is needed to obtain the desired 15,750 respondents from the RDD frame.

5. INTEGRATION OF THE TWO FRAMES

5.1 Field Sample Control

The regional component of the CCHS collects data on a monthly basis and hence requires the preparation of monthly samples for the field collection unit. Each month, the process begins with the sample selection of dwellings from the area frame. The sample sizes are adjusted to account for vacant dwellings (~12%) and the anticipated household nonresponse (10% for the face-to-face interviews). Depending on the availability of the area frame sample for a particular month versus the required monthly sample sizes for each HR, the RDD sample sizes are derived for those HRs where needed. These RDD sample requirements are adjusted to account for the HR-level hit rates and the anticipated household nonresponse (15% for the telephone interviews).

Because of the magnitude of the sample and the use of two overlapping frames, an unduplication process has been put into place. Telephone numbers are collected from the area frame respondents, and are added to an exclusion file so that they can be excluded from future RDD monthly selection. Similarly, RDD respondents are asked to provide their addresses so that they can be excluded from future area frame monthly selection.

5.2 Weighting and Estimation

The development of the weighting strategy for the regional component is currently underway. The plans discussed to date are to have two separate weighting systems, each with its own set of weight adjustments. The area frame adjustments would include, among other things, adjustments for cluster growth, stabilisation, and household and person level nonresponse. The RDD weighting system would include adjustments for no phone lines, multiple phone lines, as well as household and person level nonresponse. The two weighting systems would be integrated using an integration method that would take into account the design effect and the effective sample sizes of each of the designs. The integrated weights would be calibrated using a one-dimensional poststratification adjustment of ten age/sex poststrata within each health region.

6. FUTURE DEVELOPMENTS

The main objective of the provincial component of the CCHS, scheduled to begin collection in January 2002, is to produce cross-sectional estimates on the different aspects of mental health and well-being of Canadians, from a sample of 30,000 respondents. The survey will collect data on both positive and negative factors affecting mental health, the utilization of mental health care services, social impacts, and the costs associated with mental health. This will be rounded out with data collection on a number of social and demographic characteristics.

Although the budget for future cycles of the CCHS is still under review, plans call for a repetition of the CCHS biennial cycle using a similar approach: a regional component with a large sample size in the calendar year 2003 followed by a provincial component in 2004.

The area frame, as designed for the LFS, is an excellent frame that suits CCHS needs in many ways, and it will likely be the primary frame for the next regional component. Not only does it cover the same target population of households, but the updating and verification process of the frame, already in place, is a definite asset and falls within the operational and budget constraints of the CCHS. RDD will also likely be used, unless modifications are made either to the stratification of the area frame or to the provincial HR delineations, to make them more compatible with each other.

There are many questions and concerns about the future of RDD sampling. The increasing popularity of technologies such as Call Screening and Call Display may effect response rates of telephone surveys.

Another issue is that of telephone number portability. As it increases it will have a great effect on the CCHS design because geography would no longer be linked to telephone numbers. A third major issue is the use of cellular phones as the principal phone, or only phone, which affects RDD surveys in two ways. First, it reduces the ability to impute the geographic location of a phone number. Second, coverage is lost since cellular phones are presently excluded. If cellular phone usage as the only phone increases to a significant level, RDD surveys would be compromised.

REFERENCES

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 section on Survey Research Methods. American Statistical Association. To be published.

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

Paton, D.A. and Dolson, D. (1998). Random Digit Dialing at Statistics Canada. Statistics Canada. Internal document.

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