Sample Design of the Canadian Health Measures Survey

Rebecca Morrison[1] and Suzelle Giroux[2]

ABSTRACT

Statistics Canada, in partnership with Health Canada and the Public Health Agency of Canada began the collection of the Canadian Health Measures Survey (CHMS) in the spring of 2007. The CHMS aims to overcome limitations of self-reported data by directly measuring health indicators, and physical activity and fitness levels from a nationally representative sample of 5,000 Canadians, aged 6 to 79. A multi-stage sample design was developed to meet the objectives and logistics of the survey. The presentation will provide an overview of the CHMS with emphasis on the sampling plan.

KEY WORDS: health survey; direct measures; multi-stage sampling

RÉSUMÉ

Statistique Canada, en partenariat avec Santé Canada et l'Agence de santé publique du Canada, a débuté la collecte de l'Enquête canadienne sur les mesures de la santé (ECMS) au printemps 2007. L'ECMS a pour but de surmonter les limites des données autodéclarées en mesurant directement des indicateurs de santé, l'activité physique et le niveau de condition physique d'un échantillon national représentatif de 5 000 canadiens âgés de 6 à 79 ans. Un plan d'échantillonnage à plusieurs degrés a été développé pour répondre aux objectifs et à la logistique de l'enquête. La présentation donnera une vue d'ensemble de l'ECMS en mettant l'accent sur le plan d'échantillonnage.

MOTS CLÉS : enquête sur la santé, mesures directes, échantillonnage à plusieurs degrés

1. INTRODUCTION

The Canadian Health Measures Survey (CHMS) collects key information relevant to the health of Canadians in the form of direct physical measurements such as blood pressure, height and weight, blood and urine specimens, and physical fitness testing. The information gathered through direct measures of health is essential to evaluate the true extent of the problems associated with such major health concerns as diabetes, obesity, hypertension, and cardiovascular disease.

All measures are taken by health professionals in a specially designed mobile clinic. Blood and urine samples are collected to measure diabetes, cardiovascular health, nutritional status, infectious disease markers, and exposure to environmental contaminants, such as lead.

This innovative survey presents many challenges, notably the logistics of physical measures data collection, the storage of bio-specimens, the maintenance and administration of the mobile clinic, and the privacy and ethics issues surrounding such information. In addition, costs that were higher than expected were coupled with limited resources. Tremblay provides an exhaustive list of the challenges facing the CHMS (publication forthcoming).

The purpose of this paper is to provide an overview of the CHMS with emphasis on the chosen sample design, which is driven by the above mentioned challenges. Section 2 provides a brief description of the survey, including objectives, data collection and dissemination. Section 3 describes the sample design, namely the multi-stage approach chosen for the CHMS to address the objectives of the survey while respecting any constraints arising from costs and collection logistics.

1 Rebecca Morrison, Statistics Canada, Canada, K1A 0T6, rebecca.morrison@statcan.ca
2 Suzelle Giroux, Statistics Canada, Canada, K1A 0T6, suzelle.giroux@statcan.ca

Finally, section 4 outlines the future work, such as non-response weight adjustments and possible variance estimation methods. The unique challenges associated with collecting direct measures are also discussed throughout the paper.

2. OVERVIEW OF THE SURVEY

The target population of the CHMS includes all persons, aged 6 to 79, residing in private dwellings. Excluded from the target population are populations on Indian Reserves, Canadian Forces Bases and members of the Canadian Forces, residents of institutions, and residents in some remote areas. The survey is designed to produce reliable national estimates for 10 groups defined by age and sex. The age groups are: 6 to 11, 12 to 19, 20 to 39, 40 to 59, and 60 to 79.

The main objectives are to:

- estimate the number of people with selected health conditions, characteristics and exposures
- estimate the distribution of selected diseases, risk factors and protective characteristics
- determine relationships among risk factors, personal characteristics, and health status
- collect information that cannot be gathered through self reported interviews
- assess the validity of prevalence estimates based on self- and proxy-reported information.

2.1 Content

There are two main components to the CHMS: an in-home interview and a clinic visit. A health questionnaire is administered during the in-home interview. The health questionnaire gathers information related to nutrition, smoking habits, alcohol use, medical history, current health status, sexual behaviour, lifestyle, physical fitness, as well as demographic and socioeconomic variables. Upon completion of the in-home interview, respondents are invited to visit the CHMS mobile clinic. The clinic portion includes physical measurements, such as height and weight, tests of physical fitness, flexibility and muscular strength, a test lung capacity and power, and an oral health examination. Blood and urine samples are collected to measure diabetes, cardiovascular health, nutritional status, infectious disease markers, and exposure to environmental contaminants, such as lead. Respondents also receive a report detailing the results of the physical measures and their travel expenses are reimbursed at the end of the clinic visit. Later, respondents receive the results of their laboratory tests by mail. Data collection began in the spring of 2007 and will continue until early 2009.

2.2 Data Dissemination

Data will be released in a staggered fashion beginning sometime in 2010. The first release will include data from the in-home interview and the clinic visit; the next release will include results of laboratory tests applied to the whole sample, followed by a release of results of laboratory tests applied only to the sub-samples. The first articles using CHMS data will include normative data tables on such topics as respiratory health, kidney function, and environmental exposure, and will be published in Statistics Canada's *The Daily*.

3. SAMPLE DESIGN

The CHMS uses a multi-stage sample design. The first stage is the selection of collection sites, the second stage is the selection of dwellings, and the third stage is the selection of people.

3.1 First stage: frame and sample

With a portion of the data collection taking place in the mobile clinic and to minimize travelling by the respondents, collection sites had to be defined. The CHMS used the Labour Force Survey (LFS) area frame to create collection sites. Information regarding dwelling composition, obtained from the 2006 Census of Population, was used to create list frames within selected collection sites.

Travel distance for respondents and population density factored in the construction of the sites. The LFS area frame was used with LFS clusters grouped to construct CHMS collection sites. CHMS collection sites require a minimum

population density of 10, 000 people and a maximum travel distance of 50 km from the site boundary to the site centre in urban sites and 100 km from the site boundary to the site centre in rural sites. Known geographical boundaries, province and Census Metropolitan Areas, were also respected when constructing sites.

A total of 257 sites were constructed. Whenever possible, sites were regrouped to avoid exclusions due to insufficient population density. That is, a site of smaller population would be grouped with a neighbouring site provided the new distance from site boundary to centre, for the new combined site, was still reasonable. Larger sites, those with distances exceeding the 50 to 100 km constraint, were divided provided the population in the resulting sites was still sufficient. Some sites were excluded because of insufficient population density or because they were part of high vacancy or remote areas. These exclusions account for approximately 3.7% of the target population.

Although only national level estimates are required for the survey, the sample allocation is done by region to ensure all regions are present in the sample. The stratification is comprised of five regions: the Atlantic Provinces, Quebec, Ontario, the Prairies, and British Columbia. The sample of collection sites is allocated proportionally to the size of the population, according the 2001 Census, in each region.

Within each stratum, the collection sites are ordered by Census Metropolitan Areas, and by ascending population size. Sites are randomly selected systematically with probability proportional to the size of the population. Even though more collection sites are desirable, and are recommended, high costs and data collection logistics resulted in only 15 collection sites being selected. For each of the 15 selected sites, the mobile clinic remains in a fixed location for approximately six weeks of field collection.

3.2 Second stage

The second stage is the selection of dwellings. A list of dwelling addresses, along with the age and sex of the residents of each dwelling, is obtained from the 2006 Census for each of the 15 selected sites. For the purpose of stratification, the age of dwellings members is calculated as of the first day of collection in a site. Dwellings are assigned to six strata based on the age profile of its residents. Dwellings are stratified in a hierarchical fashion from the most difficult to the least difficult to target age groups. First, all dwellings with at least one person aged 6 to 11 are grouped into the 6- to 11-year-old stratum. Next, the remaining dwellings with at least one person aged 12 to 19 are grouped into the 12- to 19-year-old stratum. This continues in the same fashion for the 60- to 79-year-olds, 20- to 39-year-olds, and 40- to 59-year-olds respectively. Finally, all other dwellings, that is those without a person aged 6 to 79 and those that were vacant at the time of the 2006 Census, are grouped into the last stratum, the out-of-scope stratum. The allocation of dwellings is done by simulation in combination with expected vacancy and response rates. The dwellings are selected using stratified simple random sampling.

3.3 Third stage

The third stage is the selection of people. CHMS interviewers visit all selected dwellings. Wherever possible, a list of all dwelling members is obtained, along with their age and sex, and one or two respondents are randomly selected. Whenever a dwelling contains at least one person aged 6 to 11, two people are randomly selected: one aged 6 to 11 and the other aged 12 to 79. Naturally, if there is more than one person aged 6 to 11, only one among them is randomly selected. If there are no 6- to 11-year-olds present in the dwelling, then only one person aged 12 to 79 is selected.

If there are 6- to 11-year-olds present in the dwelling, one will always be selected. Furthermore, individual selection probabilities within a dwelling may vary as the remaining age groups are favoured for selection in their respective strata. Dwellings grouped in the 20- to 39-year-old stratum, for instance, favour the selection of 20- to 39-year-olds by giving twice the chance of selection to persons aged 20 to 39 over all other age groups. Since the 12- to 19-year-olds are hard to target, they are not only favoured for selection in their respective strata, they are also favoured for selection in the 6- to 11-year-old stratum (for the second person selected) and the out-of-scope stratum.

If a person is selected in an age group other than the one targeted in a particular stratum, larger, and in some cases extreme, person selection weights can result. For example, if a dwelling contains six people and five of them are in the same age group and are all given twice the chance of selection over the sixth person, and the sixth person is selected then the person selection weight is 11. If one of the other five is selected, the person selection weight is only 5.5. In order to

reduce extreme person selection weights, caused by the favouring of certain age groups, an equal probability of selection for all age groups is used whenever the number of people in the dwelling is above a certain threshold.

3.4 Sample size

The CHMS aims to produce reliable national level estimates for each of the 10 age and sex groups. The sample size is based on the goal of estimating a prevalence rate of 10% with a coefficient of variation of 16.5% and an assumed design effect of 1.5. This leads to approximately 500 respondents in each of the 10 age and sex groups for a total of 5000 respondents. The design effect was assumed to be 1.5 based on a sample of 5000 distributed equally in 30 sites. The number of sites was reduced to 15 because of costs, and logistical and operational constraints, but it was not possible to increase the sample size, so it is conceivable that the design effect is greater than 1.5 and that the desired level of precision may not be achieved.

The number of dwellings selected within each stratum is determined based on the requirement of an equal number of respondents (500) in each age and sex group completing both components of the survey: in-home interview and clinic visit. This raw sample of dwellings is inflated to compensate for vacancies and non-response accordingly. Non-response occurs at different levels: selected dwellings that do not respond; selected persons who do not complete the in-home interview; and selected respondents who complete the in-home interview but do not complete the clinic visit. Expected vacancy and non-response rates are used to determine the number of dwellings initially selected. As it is crucial to achieve the desired number of respondents, additional dwellings are randomly sampled in advance and left as a back-up, to be released if need be for field collection.

3.5 Sub-samples

There are a few sub-samples that are randomly assigned during the household interview intended for specific analytical requirements. First, all respondents visiting the clinic are divided randomly into two groups: fasted and non-fasted. To meet fasting requirements, respondents are asked not to eat or drink anything other than water 12 hours before their clinic visit. The difference between the fasted and non-fasted respondents is the tests performed on their urine and blood. Certain biological tests are impaired by the consumption of food or drink.

The CHMS is also measuring exposure to environmental contaminants on a sub-sample of respondents. There are seven environmental measures, such as inorganic mercury, for which respondents are randomly and independently selected. These tests are quite expensive, and this is the reason only a sub-sample will have them performed. Certain measures require that only one respondent in a dwelling be selected because it is believed that residents of the same dwelling have similar exposure levels, and the limited funding would be better used on a respondent from a different dwelling, rather than a second respondent selected from the same dwelling. In the case where two respondents are selected from one dwelling, a collocated selection method was used to select the sub-samples to minimise the occurrence of both respondents being selected.

3.6 Repeated measures

Groves defines measurement error as inaccuracies in response recorded on the survey instruments. These can be caused by the respondent, the interviewer, the questionnaire, the data collection method, or the measuring tool (1989). This section deals specifically with the measurement errors arising from collecting direct measures of health. When data collection is done poorly, the variance may be increased and bias introduced. Improper techniques used when taking measures, improper calibration of instruments, or uncontrolled conditions within a laboratory may result in observations that are not reflective of the true value. There are tools in place to ensure these types of errors are kept to a minimum: training and re-training of health measures specialists (HMS), calibration of measurement equipment, and proper handling and storage of biological specimens. The number of HMS is kept small (four regular HMS and two senior HMS) to minimize the variability associated with more than one HMS collecting physical measures. To ascertain the reproducibility of the physical measurements, three physical measures are repeated: anthropometry, grip strength, and sit and reach. Each of these measures is repeated on a non-overlapping randomly selected sub-sample of 10% of the clinic respondents. The goal is to end up with an equal mix of repeated measures performed by the same HMS and by a different HMS to measure reproducibility of the measures. Respondents are randomly assigned to repeat one physical measure; however, the HMS is not randomly assigned to perform the repeated measure. Assignment of HMS is done at

the clinic based on availability, resulting in a disproportionate amount of physical measures being repeated by the same HMS as opposed to a different HMS. As the design of the repeated measures is not controlled for, i.e., the HMS performing the repeated measures are not assigned at random, the data analysis is limited. For this reason, the plan is to evaluate consistency of measurement only, and not to correct variance estimates based on findings from the repeated measures.

4. FUTURE WORK

4.1 Non-response

Early results from the first five sites of collection indicate certain age and sex groups have lower response rates to the in-home interview and to the clinic visit, as well as to the blood and urine specimen collection. There are corrective measures in place to ensure sufficient respondents in each age group complete both components of the survey. Response rates at various levels are monitored on a site by site basis and modifications can be made 1) to the number of dwellings initially selected in each stratum, 2) to the person selection probabilities, and 3) to the sub-sampling rates (described in section 3.5). In addition, extra dwellings are randomly selected at each site and left as a back-up to be released for field collection should the number of respondents visiting the clinic be too low. Having this back-up of dwellings allows a less conservative approach in determining the initial sample size of dwellings. Beyond any modifications made to the sample itself, a non-response bias weight adjustment strategy, using auxiliary information, is applied at the different levels of non-response.

Of the 2759 in-scope dwellings selected for the first five sites of the CHMS, 1921 provided the dwelling composition, for a dwelling response rate of 70% (not shown). There are 2355 selected respondents coming from the 1921 responding dwellings, and 2035 selected respondents completing the in-home interview for an in-home interview completion rate of 86%. Of the 2035 respondents completing the in-home interview, 1746 visited the clinic for a clinic visit completion rate of 86%. Table 1 presents the overall response rates for the first five sites.

Table 1 – CHMS Response Rates Based on First Five Sites

Age group	Selected respondents	Completing in-home interview	In-home interview completion rate	Completing clinic visit	Clinic visit completion rate	Combined rate
6-11	433	375	87%	339	90%	78%
12-19	363	314	87%	278	89%	77%
20-39	508	442	87%	361	82%	71%
40-59	605	510	84%	434	85%	72%
60-79	446	394	88%	334	85%	75%
Total	2355	2035	86%	1746	86%	74%

4.2 Estimation

In this section sampling weight refers to the final adjusted weight used to make inferences about the population. Several sampling weights will be produced to meet the estimation requirements of the CHMS. There will be sampling weights calculated and assigned to all clinic respondents, so that inferences about the population are made based on the respondents visiting the clinic. In addition, a set of sampling weights will be produced for the clinic respondents providing biological specimens, for the fasted clinic respondents, and for each of the seven independent environmental sub-samples.

Weighting classes based on Response Homogeneity groups (RHG) will be created to lessen the bias associated with non-response by grouping similar respondents and non-respondents. For the CHMS, logistic regression and the segmentation model method will be considered for creating the weighting classes.

The variance estimation for the CHMS is a challenge in itself. With a sample of approximately 5000, less for the sub-samples, and a small number (15) of primary sampling units (PSUs) stratified in five regions; careful consideration will

have to be given to the analytical needs when developing a method. Users of Statistics Canada health data are accustomed to variance calculation using the bootstrap method, and it is possible that this is the preferred route for the CHMS. An advantage of this method is that a bootstrap program exists that may be provided to the users with the data release.

A replicate-based variance estimation method is commonly used to estimate the variance of an estimate under a complex sample design. As Sharon Lohr describes, a replicate-based method of variance estimation is appropriate under a first stage sampling with replacement design (298). The method may still be used when the sampling is performed without replacement, as is the case for the CHMS, but this tends to result in an overestimate of the variance. Employing the balance repeated replication variance estimation method would require creating pseudo strata. These new pseudo strata would contain two and only two PSUs. As the CHMS has 15 PSUs, splitting or regrouping PSUs would also be necessary. The jackknife method is not considered promising as percentile estimates are required for the CHMS. Calculating the variance directly is a possibility, although the portion of the estimate coming from the person selection will be complex since the CHMS is selecting more than one respondent some of the time. Performing a random grouping may also be an option as this method is easy to compute and performs well for percentiles. The draw back of the random group method is that the number of groups constructed is limited by the survey design (Lohr, 297). Whatever the choice of variance estimation method, it will be one that meets the estimation objectives of the survey as closely as possible while taking into account the sample design and addressing the needs of the data users.

5. CONCLUSIONS

There were many challenges during the development of the survey, and the current challenges relate to the collection of physical measures. Initially, it was unclear whether Canadians would participate in such a survey; however, early data collection results are promising. After the first five sites participation levels are higher than expected. In the next years, the focus will be devoted to the task of dissemination. More challenges are expected in preparing the data files, developing the weighting methodology, and determining the variance calculations for the CHMS.

6. ACKNOWLEDGEMENTS

The authors wish to thank the CHMS team, both past and present, for their dedication. Special thanks go to Nicolas Lavigne for his contribution to the CHMS sample design, and to John Rowland and Jack Singleton for their valuable comments in reviewing this paper.

REFERENCES

Groves, Robert M. (1989), *Survey Errors and Survey Costs*, New York: John Wiley & Sons.

Lohr, Sharon L. (1999), *Sampling: Design and Analysis*. Pacific Grove, Ca: Duxbury Press.

Tremblay, M. S. (2006), *Learning the Unique and Peculiar Challenges of Direct Health Measures Surveys: The Canadian Experience*. Paper presented at the 2006 Statistics Canada Symposium: Methodological Issues in Measuring Population Health. To be published.