# THE METHODOLOGICAL CHALLENGES OF THE 2009 SURVEY ON LIVING WITH CHRONIC DISEASES IN CANADA (SLCDC)

Mamadou S. Diallo[1], Marie-Claude Duval[2], Steven Thomas[3]

## ABSTRACT

The purpose of the Survey on Living with Chronic Diseases in Canada (SLCDC) is to provide information on the impact that chronic disease has on individuals, as well as how people with chronic disease manage their health condition. In 2009, the survey covered arthritis and hypertension, and followed-up on a sub-sample of respondents from the 2008 Canadian Community Health Survey (CCHS). This paper will present an overview of the 2009 SLCDC survey and some of its challenges; one of them being the weighting for the higher than expected out of scope rates as well as the unresolved cases. Unresolved cases are those where it could not be determined if they were out-of-scope or not. Therefore, modeling was used to predict the out-of-scope sub-population among the unresolved cases. The paper will also cover the sampling process, as well as the weighting and the variance estimation method used. Using replicate methods like bootstrap, in the context of a two-phase complex survey, raises some theoretical questions. Some of them will be answered in this article, but the evaluation of the variance estimation method used is still ongoing.

KEY WORDS: Bootstrap, Nonresponse, Sub-sample, Weighting.

## RÉSUMÉ

Le but de l'enquête sur les personnes ayant une maladie chronique au Canada (EPMCC) est de fournir de l'information sur l'impact des maladies chroniques sur les individus et sur la façon dont ces derniers gèrent leur condition. En 2009, l'enquête s'est intéressée à l'arthrite et l'hypertension en faisant un suivi d'un sous-échantillon de répondants de l'enquête sur la santé dans les collectivités canadiennes (ESCC) de 2008. Le présent article discute de l'EPMCC 2009, ainsi que certains des défis méthodologiques survenus, comme l'ajustement de la non-réponse pour tenir compte d'un taux plus élevé que prévu d'unités hors cibles et de la présence de cas non résolus. A cause de la présence de cas non résolus, la modélisation a été utilisée pour prédire la sous-population des cas d'hors cibles parmi les non résolus. Le processus d'échantillonnage, de pondération et la méthode d'estimation de la variance utilisés seront également présentés. L'utilisation des méthodes de réplication, comme le bootstrap, dans le cas des enquêtes complexes à deux phases soulève quelques préoccupations théoriques. Certaines réponses sont fournies dans cet article. Cependant l'évaluation de la méthode d'estimation de la variance est toujours en cours.

MOTS CLÉS : Bootstrap;, non-réponse; pondération; sous-échantillon,.

## 1. INTRODUCTION

The purpose of the Survey on Living with Chronic Diseases in Canada (SLCDC) is to provide information related to the experiences of persons living with chronic diseases, including diagnosis of a chronic health condition, care received from health professionals, medication use and self-management of their condition. The SLCDC is a cross-sectional survey and a follow-up to the Canadian Community Health Survey (CCHS). For more information on the CCHS one can consult the CCHS User Guide or Sarafin and al. (2007). In other words, the SLCDC is a two-phase survey in which CCHS is the first phase. Therefore, the respondents of the CCHS form the second phase population from which the SLCDC sample is selected. The SLCDC takes place every two years and each year two or more chronic diseases are covered. Arthritis and hypertension were covered in 2009.

The purpose of this paper is to give an overview of the survey and to describe the methodology used. In Section 2, an explanation of the 2009 SLCDC sampling process in the context of a two-phase process is provided. For about 20% of the

[1] Mamadou S. Diallo, Statistics Canada, Tunney's Pasture, R. H. Coats Building, Ottawa, ON, K1A 0T6, Canada, MamadouSaliou.Diallo@statcan.gc.ca
[2] Marie-Claude Duval, Statistics Canada, Tunney's Pasture, R. H. Coats Building, Ottawa, ON, K1A 0T6, Canada, Marie-Claude.Duval@statcan.gc.ca
[3] Steven Thomas, Statistics Canada, Tunney's Pasture, R. H. Coats Building, Ottawa, ON, K1A 0T6, Canada, Steven.Thomas@statcan.gc.ca

sample, called unresolved cases, it was not possible to determine whether or not the selected person was in the target population. Several methods were evaluated to address this issue. In Section 3, we will present the 2009 SLCDC weighting method used. Section 4 presents the bootstrap variance estimation method used in the context of two-phase sampling. This method raises some challenges and still needs more evaluation in order to be fully understood.

## 2. SAMPLING

### 2.1 Sampling Frame

The 2009 SLCDC used the 2008 CCHS respondents to select its sample. Therefore, it is called a two-phase design, in which the first phase is the CCHS sample and the second phase is the SLCDC sample. The second phase survey population can be defined as the CCHS 2008 respondents living with arthritis or hypertension aged 20 years old or older and living in the 10 provinces. The territories were not covered in the 2009 SLCDC. In addition, all the exclusions applied to the CCHS also apply to the SLCDC. These exclusions are persons living on Indian Reserves or Crown lands, full-time members of the Canadian Forces, residents of institutions and certain remote regions.

Extra exclusions were made for practical reasons. People with invalid phone numbers were excluded since it was impossible to contact them given that the SLCDC was a telephone survey. People who did not agree to share or link their CCHS 2008 information were also excluded. This is because the intent of the SLCDC was to link the SLCDC survey responses with the CCHS 2008. This linked-share survey data file, containing only respondents who gave their permission to share and link, will then be provided to the survey share partners such as the survey sponsor, the Public Health Agency of Canada (PHAC). People for whom their CCHS interview was done by proxy were excluded since the SLCDC questionnaire could not be answered by proxy. These were exclusions from the sample population rather than the survey's target population and were treated in the weighting (see section 3.1). Table 1 gives the 2nd phase survey population counts, the exclusions counts and the remaining counts available after all the exclusions. The 2nd phase survey population is the respondents from the CCHS that are in our population of interest. Note that the sampling will be performed from the units remaining after the exclusions for practical reasons.

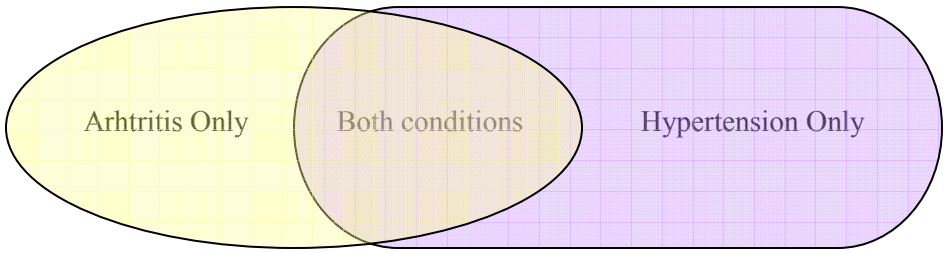
**Table 1: Exclusions counts and percentages for each condition**

| Condition | 2nd Phase Survey Population | Exclusions for practical reasons | After exclusions |
|---|---|---|---|
| Arthritis | 13549 (100%) | 2581 (19.05%) | 10968 (80.95%) |
| Hypertension | 17437 (100%) | 3224 (18.49%) | 14213 (81.51%) |
| Both conditions | 24048 (100%) | 4448 (18.50%) | 19600 (81.50%) |

### 2.2 Sampling Strategy

In order to produce reliable estimates at the national level by age groups by sex, the remaining people from the CCHS sample were stratified based on the age groups of interest: 20 to 44 years old, 45 to 64 years old, 65 to 74 years old and 75 years and older for each of the chronic condition. The sample size needed to estimate a prevalence of 10%, with a CV of 16.5% based on the assumptions that the design effect was 2.8 and an overall response rate of 70% is estimated at 1,324 by stratum (age groups by sex) and condition. One constraint was that only one questionnaire could be administered per unit for a response burden issue. This means that people with the two conditions will get only one questionnaire. For that reason, each stratum was partitioned into three parts, which are the people with arthritis, those with hypertension, and the people with both conditions. The figure 1 shows the overlap between the two conditions.

Figure 1: Overlap oh the two conditions




In the case that the number of units available was larger than the requested sample by condition, a sample allocation was done proportional to the second-phase population for each part. Then the sample selection was done independently from each part and condition using a systematic sampling by province, collection period[4] and age. In the case that the number of units available was smaller that the require sample by condition, all units were selected for the part 'one condition only'. For the third part where people have both conditions, two samples were selected in order that each unit receives only one questionnaire (the people from one sample for arthritis questionnaire and the people from the other sample for the hypertension questionnaire). In that case, the sample was allocated in such a way that sample sizes at the intersection level of the two conditions are proportional to the condition's sizes at the stratum level. The sample sizes by stratum by condition are given in Table 2.

**Table 2: Sample sizes by stratum and condition**

| Strata (Age group by Sex) | Arthritis | Hypertension |
|---|---|---|
| M2044[5] | 343 | 619 |
| M4564 | 1,167 | 1,324 |
| M6574 | 616 | 1,289 |
| M75p | 574 | 1,007 |
| F2044 | 503 | 844 |
| F4564 | 1,324 | 1,324 |
| F6574 | 1,224 | 1,324 |
| F75p | 1,311 | 1,324 |
| Total | 7,062 | 9,055 |

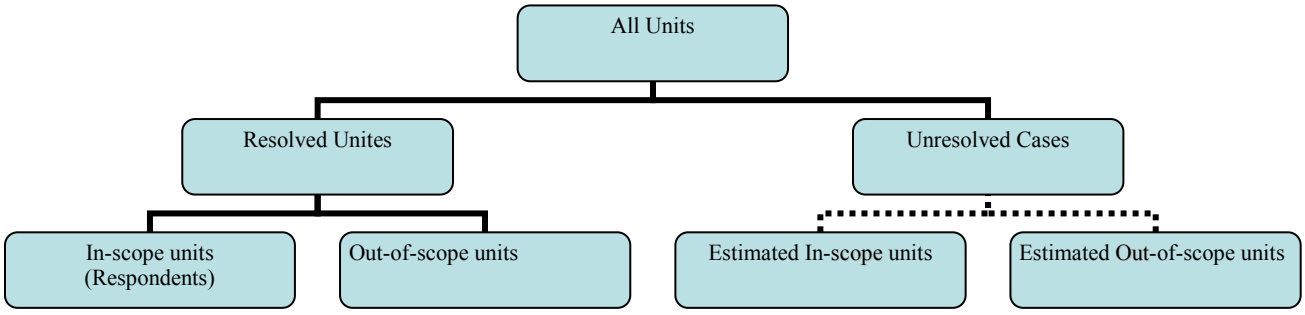
## 3. WEIGHTING

In this section, the adjustments made to the weights, to ensure that the survey respondents represent the target population, are presented. The adjustments are discussed in the same order that they were performed in the SLCDC 2009. The figure 2 shows the general classification of the units within a stratum (age group by sex) by condition. The lines indicate the classification of the units as respondents, out-of-scope or unresolved obtained before the weighting only by using the survey responses. The dotted lines indicate the modeling used to predict the probability of being in scope among the unresolved cases (see section 3.3).

All the weight adjustments took place at the age group by sex by condition level.

[4] CCHS has a continuous collection approach in which a sample is selected and collection performed every 2 months.
[5] M2044 represents males aged 20 to 44 years.

**Figure 2: Classification of units as respondent / non respondents or out-of-scope**




### 3.1 CCHS Proxy-Link-Phone Adjustment

Since those excluded people for practical reasons were still considered in the population of interest, adjustments were made to allocate their weights to the remaining CCHS respondents available for sampling. Weights adjusted for the people who refused to share their information, called share weights, were already available from the CCHS 2008. To adjust the share weights for people who did their interview by proxy, who refused to link their data and who did not have a valid phone number, a logistic modeling was used with CCHS 2008 variables as auxiliary variables to determine the probabilities of being in scope. Response Homogeneity Groups (RHGs) were formed based on similar predicted probabilities. In each group, the adjustment factor corresponded to the inverse of the weighted proportion of people who were not excluded.

#### 3.1.1 Selection Adjustment

3.1.2 As described in section 2.2, there are four different groups of people independently selected by age group by sex. These four groups are people with arthritis only, people with hypertension only, people with both conditions answering the arthritis questionnaire and people with both conditions answering the hypertension questionnaire. In each stratum the weight adjustment was done separately in each of those four groups.

### 3.2 Unresolved Adjustment

Some people declared in the 2009 SLCDC interview that they never had the condition for which they were selected to participate in the survey. These situations lead to higher than expected out-of-scope rates, 12.63 % for hypertension and 17.45 % for arthritis. Out-of-scope units are simply dropped from the weighting process whereby the counts estimated from SLCDC will be lower than those from CCHS. This created challenges in relation to the unresolved cases. The unresolved cases corresponded to people for whom it was not possible to determine whether or not they were in the target population. This mainly happened in non-completed interviews, such as refusals and non-contacts, in which the selected person failed to confirm that she or he had the condition. Due to the high percentage of respondents who declared that they never had the condition, it was not possible to assume that these unresolved cases were in-scope. Therefore, in the SLCDC, modeling was necessary in order to estimate the number of people among the unresolved cases who were not part of the target population as well as those who were part of the population and should be included in the nonresponse adjustments. The weight adjustment for unresolved cases was done in two steps. The first step is the out-of-scope adjustment and the second step is the nonresponse adjustment.

#### 3.2.1 Out-of-Scope Adjustment

A logistic regression model, predicting the probability of being in-scope was fit. The resolved cases (respondents and known out-of-scope cases) were used with their variables from the 2008 CCHS as auxiliary variables. The resulting model was then used to predict the probability of the unresolved units to be in-scope. The predicted probability of a specific unresolved unit to be in-scope was multiplied by its weight to obtain the estimated number of in-scope people in the population represented by that specific unit. This step can be called the adjustment for out-of-scope cases.

### 3.2.2 Nonresponse Adjustment

At this point, the known out-of-scope units have been dropped and the unresolved unit weights have been adjusted for out-of-scope. The unresolved units with their adjusted weights were considered to represent non respondents. The weights of the respondents were adjusted within each RHG using logistic regression modeling. The fact that all the out-of-scope cases were not known so modeling had to be done to estimate them among the unresolved cases, added variability and complexity to the weighting process. On the other hand, all the unresolved cases were kept for the nonresponse modeling with their adjusted weights. This gave more flexibility in forming the RHGs rather than the situation where part of the unresolved group was known as out-of-scope and dropped from the modeling.

## 3.3 Final (Share/Link) SLCDC Weights

About 2% of the respondents refused to share their information to other agencies outside of Statistics Canada or to link their information with administrative files. Since this number was small, it was decided to exclude them from the data file. Therefore, they were excluded from the final product and adjustments were made to allocate their weights to the remaining SLCDC respondents. The probability of agreeing to share and link the SLCDC data was predicted for each respondent and RHGs were formed based on similar predicted probabilities. In each RHG, the weights were adjusted as described in 3.1. The weight obtained for each remaining unit was the final weight for the 2009 SLCDC. These exclusions allows to work with only one file that can be sure with partners and link to administrative data within Statistics Canada.

# 4. VARIANCE ESTIMATION

## 4.1 Idependent and identically distributed (i.i.d) samples

The technique used for the variance estimation is the bootstrap method. This method is used widely by complex surveys statisticians, and, in particular, the CCHS used it to estimate the variance. In the context of a single phase sampling, theoretical results and numerous empirical studies are available to demonstrate the consistency of the method. Shao J. and Tu D. (2005) is an excellent account of the replicate methods (bootstrap, jackknife, balanced repeated replication (BRR)) in the case of samples of independent and identically distributed (i.i.d.) observations and the authors put the emphasis on the theoretical results. In the case of complex survey samples, many modified bootstrap methods were proposed to take into account the tendency of the naïve bootstrap method used in the i.i.d to be inconsistent. Among them, the CCHS used the rescaling bootstrap method proposed by Rao and Wu (1988), Wu and Yue (1992). In this method, the weights are rescaled in such a way that the variance estimators reduce to the standard ones in the special case of linear statistics.

## 4.2 SLCDC 2009 context (complex survey)

The 2009 SLCDC is a two-phase sample design and there is very little theoretical work available on the use of replication methods in the context of two-phase sampling. Kim et al. (2006) is one of the few articles that show, in some particular situations, the consistency of the replication methods, in particular, the bootstrap for a two-phase sample. The results from Kim et al. (2006) are shown in the general framework of a two-phase survey in which the first phase has a multi-sample design (stratified and clustered) and the second phase is a stratified simple random sampling. The 2009 SLCDC sample was selected by the stratified simple random method at the second phase. The 2009 SLCDC used a reweighted expansion estimator (REE) which can be defined as: $\hat{Y} = \sum_{g=1}^{G} \left( \frac{\sum_{i \in A_1} w_i x_{ig}}{\sum_{i \in A_2} w_i x_{ig}} \right) \sum_{i \in A_2} w_i x_{ig} y_i$ where $g$ indicates the second phase group, $A_1$ and $A_2$ indicate respectively the set of indices in the first phase and in the second phase. $x_{ig}$ takes value 1 if unit $i$ belongs to the $g^{th}$ group and 0 otherwise and $w_i$ is the first phase weight. In the context of the SLCDC, $w_i$ was the final weight of CCHS 2008 and $g$ could be either the stratum for the selection weight adjustment or the RHG for the nonresponse adjustments. The term in the brackets corresponds to the adjustment factor.

The SLCDC replicate weights were created from the 2008 CCHS replicates. Each of the 2008 CCHS bootstrap replicates were submitted to the same weight adjustments as mentioned before (see Section 3). Kim et al. (2006) showed that the consistency of the variance estimator remains valid with more than 2 phases. Under the assumption that the nonresponse mechanism is random, every nonresponse adjustment made can be seen as a phase. This basically means that the theoretical results are still valid despite the several adjustments made to the weights. However further evaluations are

taking place in order to fully validate the bootstrap method used. Some questions to be answered include: how do the high sampling rates for clusters affect the estimation of the variance? Should, the rescaling of the weights in the Rao-Wu method be different taking into account the second phase?

## 5. CONCLUSION

The Canadian Community Health Survey (CCHS) has been used for several follow-up surveys. Therefore, the work undertaken for the Survey on Living with Chronic Diseases in Canada (SLCDC) can benefit all of them. The 2011 cycle of the SLCDC will follow up respondents to the 2010 CCHS respondents living with asthma, chronic obstructive pulmonary disease (COPD) and diabetes. These chronic diseases have significantly smaller prevalence rates than arthritis and diabetes, which means there will be fewer units available from the CCHS. At the same time, the age groups of interest are different for each condition. Sampling methods used in 2009 will have to be reevaluated to integrate these extra challenges. The weighting process will probably be similar. However, the modifications made to the sampling process may affect the weighting process significantly. The estimation of the variance is still under evaluation. It seems that to fully understand the replication variance estimation used, it might be essential to question the first phase (CCHS) portion.

## REFERENCES


Beaumont, J.F. (2005). "On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse Through Weight Adjustment". *Survey Methodology*, **31**, 227-231.

Brisebois, F. and Thivierge, S. (2001). "The Weighting Strategy of the Canadian Community Health Survey". *2001 Proceedings of the American Statistical Association Meeting, Survey Research Methods Section*.

Kim, J. K., Navarro, A. and Fuller, W. A. (2006). "Replication Variance Esimation for Two-Phase Stratified Sampling". *Journal of the American Statistical Association*, **101**, 312-320

Rao, J.N.K., and Wu, C.F.J. (1988), Resampling inference with complex survey data. *Journal of the American Statistical Association*, **83**, 231241.

Shao, J., and Tu, D. (1995), *The Jackknife and the Bootstrap.* New York: Springer Verlag.

Sarafin, C., Thomas, S. and Simard, M. (2007). "Review of the Weighting Methodology for the Canadian Community Health Survey". 2007 *Proceedings of the Statistical Society of Canada, Survey Methods Section.*

Simard, M., Leesti, T. and Denis, J. (2003). "Tracing and Non-response Adjustment for the Longitudinal Survey of Immigrants to Canada". *2003 Proceedings of Statistics Canada Symposium.*