# Review of the Weighting Methodology for the Canadian Community Health Survey

Steven Thomas, Senior Methodologist, Household Survey Methods Division
Cathlin Sarafin, Methodologist, Household Survey Methods Division
Michelle Simard, Chief, Household Survey Methods Division

Key Words: Weighting, Cumulative Samples, Nonresponse, Integration

## 1 Introduction

The Canadian Community Health Survey (CCHS) was originally designed as two distinct surveys that would alternate on an annual basis. The first survey, or the regional component, was designed to collect general health information from a sample large enough to provide estimates for more than 100 health regions (HR) in Canada. This requires a sample of more than 130,000 respondents. To date, there have been three regional component surveys conducted in the years 2001, 2003 and 2005. The second survey, or the provincial component, was designed to focus on a specific health topic and to collect data from a smaller sample in order to provide estimates at the provincial level. This generally requires a sample of over 30,000 respondents. To date, two of these provincial component surveys have been conducted in the years 2002 (mental health) and 2004 (nutrition).

Even with the large sample size, these surveys are unable to meet the increasing demands of the health data users. These demands include having more varied subject matter, having a greater level of detail in the estimates produced and increasing the timeliness of the data released. In order to meet these demands, it was felt that a redesign of the regional component would improve the efficiency and flexibility of the survey, while the provincial component methodology would continue to be designed in relation to the subject matter being studied. Therefore, for the remainder of the paper, only the CCHS regional component will be discussed.

Under the redesign, the option of continuous collection has been implemented, where some portion of the sample will always be in collection rather than the original process where data was collected for a cycle of the CCHS and then collection would stop until the next sample was ready for collection. As well, both the questionnaire content and the data dissemination approach have been modified. Changes to the questionnaire and the collection of the sample were implemented starting in January 2007. The sampling design has remained relatively unchanged from the previous design and the weighting strategy is now under review to ensure that the methodology is able to reflect the changes that come with the redesign. For more information on the redesign refer to Bèland et al (2005).

This paper describes the issues related to the weighting of the CCHS and the improvements to the existing methods will be discussed. Section 2 provides some background on the complex, multi-frame design of the regional component of the CCHS. The weighting strategy is discussed in section 3, along with all of the proposed changes. Section 4 describes the outstanding issues facing the weighting methodology, and is followed by a conclusion.

## 2 The Redesigned CCHS

The target population for the Canadian Community Health Survey is all persons aged 12 years or older who are living in private dwellings in the ten provinces and three territories. Persons living on Indian Reserves or Crown lands, clientele of institutions, full-time members of the Canadian Forces and residents of certain remote regions are excluded from the survey.

To survey the target population, the CCHS uses three separate frames: an area frame, a telephone list frame and a random digit-dialing (RDD) frame. In most health regions, the area frame is used in conjunction with either the list frame or RDD frame, with two equally sized samples selected respectively from each frame. The area frame sample design is similar to that of the Canadian Labour Force Survey (LFS) (Gambino et al, 1998). For the telephone list frame, a simple random sample of telephone numbers is selected from publicly available lists from across the country. Finally, the RDD frame is used in certain remote areas where the quality of the telephone list frame is considered poor. Sampling for the RDD frame uses the Elimination of Non-Working Banks (ENWB) method that is used by the General Social Survey (GSS) (Norris and Paton, 1991).

### 2.1 Continuous Collection

Users familiar with the CCHS will note that the information presented thus far is common to previous CCHS surveys. The main difference in the methodology is the implementation of continuous collection. In the past, a sample of 130,000 respondents was collected over a 1-year period on a biennial cycle. As of January 2007, collection is conducted continuously to obtain an annual sample of 65,000 respondents. To ensure that collection is continuous, the annual sample is further broken down into smaller sub-samples that are each representative of the population. For the area frame, the annual sample is broken down into two sub-samples that are each representative of the population. Each sub-sample is collected over a 6-month period, with one third of the sub-sample collected every two months. The annual telephone frame sample is broken down into six representative non-overlapping sub-samples that are each assigned to a 2-month collection period throughout the year. Ideally, in the future, the annual samples from both frames will be divided into six representative non-overlapping sub-samples. However, before this can be implemented, changes need to be made to the sampling design of the area frame.

With this continuous collection design, different samples, each representing a different time period, can be cumulated to represent longer periods of time and incidentally increase the precision of estimates for fixed domains of interest. This is similar to Leslie Kish's idea of a 'rolling sample' as discussed by Alexander (2002) but differs in the fact that the sample does not 'roll' over the entire population but repeatedly surveys the evolving population until the desired number of respondents is obtained. The size of the domain, the precision required, the prevalence rate being estimated, along with the subject matter, will dictate the sample size and therefore the length of the collection period required. With a small sample collected over a short period of time, estimates can be calculated for general domains on fairly prevalent characteristics that do not follow a seasonal pattern. Estimates for more detailed domains or rare characteristics will require a larger sample size and thus a longer collection period. Characteristics with seasonality will require at least one year of collection.

After six months of collection, estimates for most prevalence rates should be publishable at the national level including an age by sex breakdown. To be publishable, it is recommended to have a coefficient of variation for the estimate of less than 33%. At the provincial level, most estimates will be publishable after 6 months, while accumulations of several years of data will be

required before an age by sex breakdown will be possible for all provinces. This is mainly due to small sample sizes in the Atlantic Provinces for the younger age groups. Estimates at the Health region level are generally publishable after 24 months of collection and several years of data will be required before breakdowns by age and sex are possible for all HRs. See Table 1 for more details. Note that this does not mean that some rates will not be publishable after shorter periods of time. Problems usually occur in the Atlantic Provinces for the younger age groups where fewer units have been allocated. In larger domains, with larger allocated sample sizes, estimates should be publishable after shorter time periods.

**Table 1: Number of Months to Calculate Publishable Estimate**

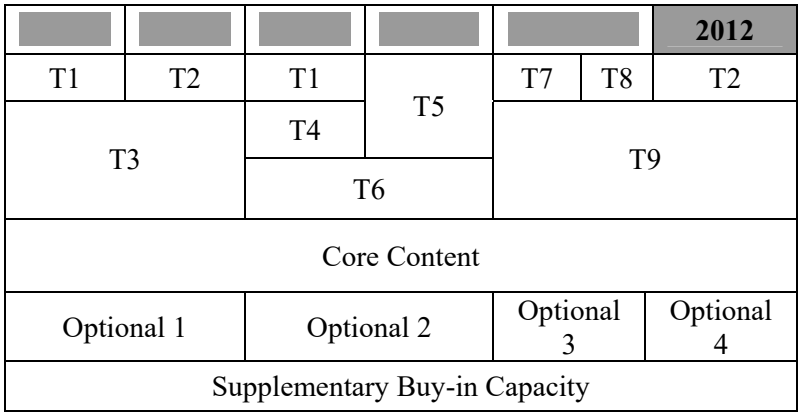
| | Prevalence Rate | | | |
|---|---|---|---|---|
| | | | | 50% |
| **Canada** | 6 | 6 | 6 | 6 |
| ***Age*** | 6 | 6 | 6 | 6 |
| ***Sex*** | 6 | 6 | 6 | 6 |
| ***Age by Sex*** | 6 | 6 | 6 | 6 |
| **Province** | 6 | 6 | 6 | 6 |
| ***Age*** | 24 | 12 | 6 | 6 |
| ***Sex*** | 12 | 6 | 6 | 6 |
| ***Age by Sex*** | >24 | 24 | 12 | 6 |
| **HR** | 24 | 12 | 6 | 6 |
| ***Age*** | >24 | >24 | 24 | 12 |
| ***Sex*** | >24 | 24 | 12 | 6 |
| ***Age by Sex*** | >24 | >24 | >24 | 24 |

### 2.2 Content

Starting in 2007, content is collected in the form of *core, theme*, and *optional* (see Figure 1). Core content is the general health information that is collected continually. Theme content, denoted by T, is more specific health information that will be collected over different time periods depending on the subject matter and precision required. Optional content is the content chosen by health regions that may be collected over a 1-year period but will generally need two years of collection in order to have a large enough sample for detailed estimation at the HR level. Finally, there is a supplementary buy-in capacity built into the survey that allows for flexibility in the collection of data on emerging issues.

For each block of content, the sample will have to be representative of the population and enough respondents will have to be collected so that the required precision is achieved. In the previous section, it was stated that the area frame required six months of collection before it was representative of the population. For that reason, the current minimum time for content to be in collection is six months for a total respondent sample size of 32,500.

**Figure 1: Content Structure for Continuous Collection**




### 2.3 Dissemination Strategy

All weights now need to be produced in a timelier manner in order to meet the demands of the new dissemination strategy. The current dissemination plan is to release annual files for 2007 and 2008, as well as a cumulative 2-year file at the end of 2008. The main reason for this strategy is that it coincides with the content structure. That is, looking at Figure 1, there is theme content specific to 2007 and 2008, as well as content that spans both years. This means that two weighting files will be produced almost simultaneously at the end of 2008. The annual files will be representative for the age and sex groups of interest at the health region level for a given year. The 2-year accumulation will allow for more precise estimates at the same level but for the 2007-2008 time period. Given the possibility of a 6-month content module in the future, as demonstrated under the year 2011 in Figure 1, it is also possible that some data will be disseminated after six months of collection. With the different weights that will have to be produced, it is clear that efficiencies must be implemented to allow for the weighting design to be more flexible and for results to be available on a timelier basis. It is also clear that the weighting strategy will have to consider the combining of weights that represent different time periods.

## 3 General Weighting Strategy

One of the challenges with the redesigned CCHS is to create a weighting methodology that incorporates the ideas of continuous collection, changing content and the revised dissemination strategy. At the same time, it is desirable to produce quality estimates that are relevant, accurate, timely, accessible, interpretable, and coherent with these weights. Given the dissemination strategy, it is clear that weights will have to be produced after each year of collection and again after two years of collection. In addition, weights for shorter time periods may be required and there will be the possibility of cumulating several years of data. Therefore, it is also desirable to have an efficient flexible process that can quickly create weights for varying periods of time. It is also important the weighting methodology ensures that the estimates produced for differing time periods are consistent.

### 3.1 Note on Estimation for a Period of Time

With most surveys at Statistics Canada, the population of interest is defined for a specific moment in time. The Census is a perfect example of this where the population is defined as of a certain

date and questions are generally asked in relation to that date (ex. May16, 2006). This is not the case with the redesigned CCHS where respondents are evaluated to see if they are in-scope for the survey at the time of interview. Note that the target population as stated in section 2 remains constant but it is not defined for any particular moment in time. This means that individuals who are not in-scope on the day of the interview may be in-scope later during the collection period. For example, individuals that turn twelve years of age during the reference period are not in-scope at the beginning of collection but are in-scope by the end of the period.

Note that the questions asked in the CCHS do not follow the typical cross-sectional snapshot idea either since there is no 'moment in time' concept with the questionnaire. Most questions are in relation to the respondent's current status, their lifetime status or their status during a certain time period in relation to the current date (ex. In the last 12 months). This means that the respondents and the characteristics of those respondents are representative of the population at the time of interview. The weights must be designed to reflect this evolving population and to estimate for evolving characteristics.

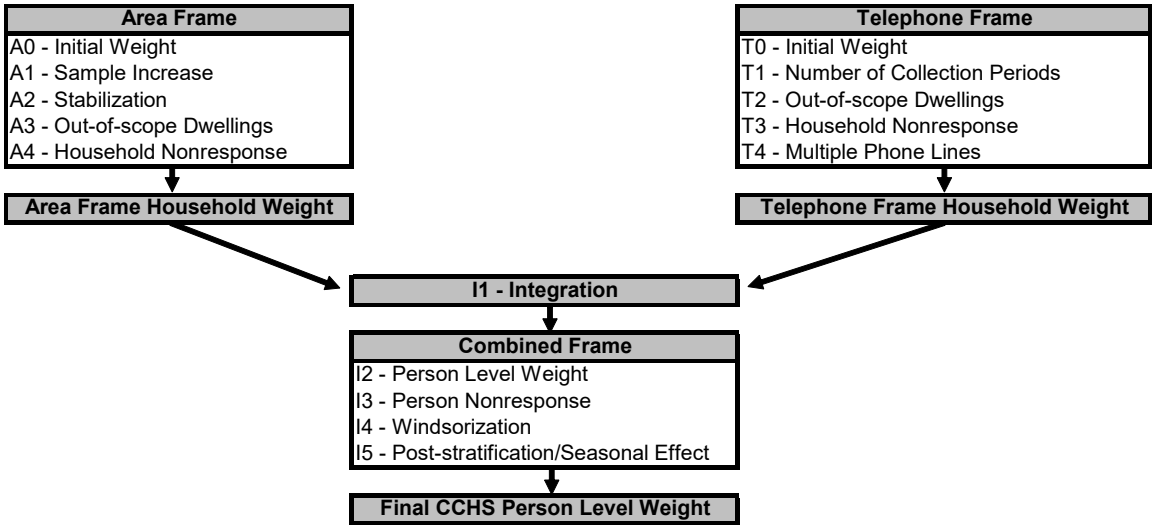
Given the evolving nature of the CCHS, the main concept that has been adopted is the idea of a period estimate, where estimates are reflective of the evolving population for a length of time. This is not a new concept for the CCHS since previous cycles of the survey were also collected over time periods. In the past, the weights were created in such a way that the different parts of the year were equally represented. More specifically, the respondents from each season represented one quarter of the year. This idea will continue with the redesigned survey. The sample will be allocated in such a way that the population will be well represented for each time period. As stated in section 2.1, the shortest time period, after which we are representative of the Canadian population, is six months with the area frame and two months with the telephone frame.

### 3.2 Weighting Overview

Like other surveys, the CCHS objective is to produce a set of final weights that yield appropriate unbiased estimates of the population of interest. Different weight adjustments are usually produced to reflect specific characteristics of the design, the propensity to respond and the population of interest. Usually, the final weights consist of three components: the sample design weights, the nonresponse adjustments applied to the design weights, and, if applicable, some calibration adjustment applied to the resulting weights after nonresponse. In the case of the CCHS, each frame has its own sampling design weights and the household nonresponse adjustments are applied to these weights. The multiple frames are then integrated and a person level nonresponse adjustment is applied. The process is finalized with the calibration adjustment.

The proposed weighting process, to be implemented for the 2007 weight file, can be seen in figure 2. Note that this differs from the process that was discussed in the past by Brisebois and Thivierge (2001). Each of the adjustments is under review to ensure that the best methodology is being used within each step (Sarafin et al, 2007). With the revised weighting process, household weights for the area frame and telephone frame are calculated separately and then integrated to have one set of weights for the entire sample. Person-level adjustments are then applied to create person-level weights, followed by a combined post-stratification / seasonal adjustment step where the weights are post-stratified to projected population counts based on the most recent census. The next sections will go into some of the weight adjustments in detail.

**Figure 2: Proposed Weighting Process for 2007**




### 3.3 Sample Design Weights

The weighting process begins on both frames with the creation of the sample design weight. For the area frame, since the sample design is based on the LFS, the weighting process begins with the weight from the LFS design. This is defined as the initial weight. Since the CCHS sample does not correspond exactly to the LFS sample, sample increase and stabilization adjustments are applied to the initial weights to obtain the actual CCHS sample design weights (combination of A0, A1 and A2 in Figure 2). For the telephone frame, simple random samples are selected which means that the initial weights are calculated simply by taking the population count and dividing it by the sample count.

In the past, the sampling design weights were calculated by treating the units as one large sample selected from one population fixed in time. However, a more accurate portrayal of the sampling process is that distinct sub-samples are collected from a continuously evolving population. This means that the weighting process will have to properly treat each of the sub-samples to reflect a fixed population for a time period. These weights can then be integrated to estimate for the evolving population for longer time periods.

### 3.4 Nonresponse Adjustments

After the sample design weights have been calculated for each collection period, the weights are adjusted to correct for total nonresponse. The nonresponse process uses a scoring method, where logistic regression is used to model the response probabilities and these probabilities are used to create the Response Homogeneity Groups (RHGs) (Simard et al, 2003). The weights of the responding units can then be adjusted to account for the nonrespondents.

One of the challenges with this methodology is that there is not a lot of information available about the nonrespondent that can be used to create the response probability model. In the past, geographic, demographic, and socio-economic information was used to model the nonresponse probabilities. The model will be improved under the redesigned methodology by adding variables that are better predictors of the nonresponse mechanism in the form of paradata.

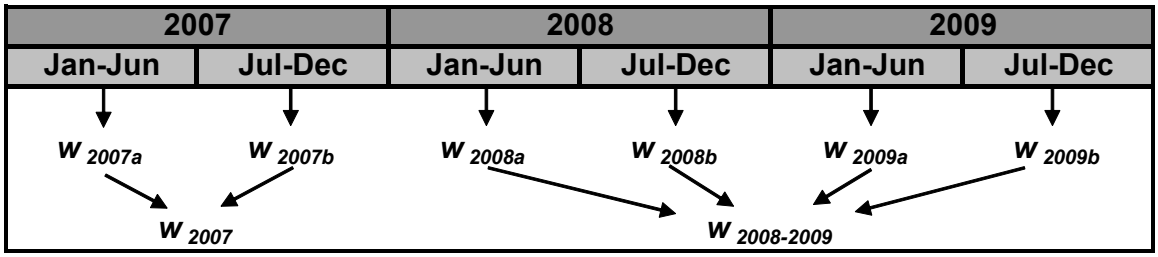
Paradata is information that is collected for every attempt made to contact a sample unit. This information can be used to derive variables such as the number of attempts, the time and day of each attempt, and whether or not an attempt was made in the evening when the probability of achieving contact is the highest. A study was done using the 2005 CCHS data which found that the addition of paradata into the model was very beneficial, especially at the household level. For the study, the nonresponse model was estimated at the health region level (sub-health region level in Québec). Without paradata, less than 22% of the HRs could be broken down into smaller RHGs, since for many health regions the original variables were not significant in the model. However, when the paradata was added, RHGs could be created in every health region since there was always at least one significant variable for the model.

### 3.5 Integration of Weights

At this point in the process, weights are available for all households for each of the frames for a specific time period. These weights must now be integrated so that the target population is represented correctly. There are different options for integrating. Weights for each frame can be integrated across time to represent the time period of interest and then the weights from both frames can be integrated to represent this time period only once. The second option is to integrate the two frames to have one set of weights representing the shorter time period followed by the integration over time. The proposed method is a combination of both ideas. As noted in section 3.3, the weights for the area frame can only be created for 6-month periods while the weights for the telephone frame can be created for 2-month periods. The first step will be to ensure that the two frames are representative of the same time period. This means that the three collection periods for the telephone frame will have to be integrated to represent the six months corresponding to the area frame. The 6-month weights from the area and telephone frames can then be integrated using dual frame techniques (Skinner and Rao, 1996). For more information on this adjustment see Sarafin et al (2007). The 6-month weights can then be integrated to represent longer time periods. This will be described in the next section.

The main advantage of such an approach is that once the weights are calculated for a 6-month time period, there should be no need to go back and revisit the weighting for that period. Weights for various longer time periods can then be created by integrating these fixed weights.

**Figure 3: Integration of Weights Over Time**




#### 3.5.1 Integration Over Time

To estimate for longer periods of time and to increase the sample size available for analysis, the sampled units representing different time periods will have to be combined. The problem with the integrated datasets is that without adjusting the weights, the population estimates are inflated by the number of periods being combined. For example, if two samples are integrated then the population total will be two times the size of the actual population. From composite estimation

techniques, suppose $k$ estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ can be calculated to estimate a population parameter $\theta$ such as a mean, proportion or total. These estimates can be integrated to calculate

$$\hat{\theta}_c = \sum_{i=1}^{k} \alpha_i \hat{\theta}_i$$

where $\sum_{i=1}^{k} \alpha_i = 1$. With this technique, if each estimate $\hat{\theta}_i$ is an unbiased estimate of $\theta$ then $\hat{\theta}_C$ will also be unbiased for all choices of $\alpha_i$. These same $\alpha_i$ can be applied directly to the original sampling weights of each period ($w_i$) to create $w_i$ *=$\alpha_i$ $w_i$.

There are several choices for values of the adjustment factors $\alpha_i$. Weights can simply be divided by the number of samples being combined, but this may be inefficient. Some methods take the sample sizes into consideration while other methods ensure that the variance of the final estimate is minimized. For more information on the different methods, see Chu, Brick and Kalton (1999) or Korn and Graubard (1995).

In order for the composite estimation techniques to be appropriate, it is necessary to assume that the estimates for each of the samples are unbiased estimates of the same parameter. As noted earlier, this is not generally the case for the CCHS. The CCHS is collecting changing characteristics from an evolving population and the weighting must take this into consideration. Therefore, the proposed method is to simply divide the weights by the number of time periods being integrated. In other words, $\alpha_i=1/k$. The main advantage of such an approach is that it leads to estimates that are easier to interpret than more efficient methods. By weighting each time period equally, the resulting estimate can be interpreted as a period estimate for the extended time period. If the different time period components were not weighted equally then the resulting estimates would be more difficult to interpret. This method will be applied when combining the weights calculated for the 6-month periods, as well as when combining the 2-month weights of the telephone frame to represent the same 6-month time period as the area frame (adjustment T1 in figure 2).

It should be noted that the estimates created with the integrated weights must be interpreted carefully. Given that the estimates for the different time periods being combined are not necessarily unbiased estimates for the same population parameter, estimates with the integrated weights will be biased for the different component time periods. That is to say that by integrating the weights, an estimate for a different population over a different time period is being created.

### 3.5.2 Calibration

The goal of calibration in the CCHS is to improve the quality of the estimates and to give coherence between the totals that are produced with the survey and the known values. Weights in the CCHS are calibrated to an average of the monthly population projections created by the Labour Statistics Division. These totals are based on the annual health region population projection totals produced by the Demography Division.

The difficulty in calibrating the CCHS comes in determining the population that is represented given the continuous collection over a time period. Averages of monthly projection estimates have been used in the past. The idea with such an approach is that any changes in the population are accounted for in the totals used. The weights allow the user to estimate for an average picture of the population over the time period. Early studies in the calibration options suggest that this

adjustment does not have a large impact on the estimated variances so the main goal of calibration is to reduce any possible bias by tying the weights to some reference distribution. Therfore, it is proposed that the weights continue to be calibrated to the average of the monthly projections for the time period of interest. With these weights, it is unlikely that inferences about the current population totals can be made unless the population counts have not changed and in general, it would be best to estimate proportions if inferences about the current population are of interest.

# 4 Challenges

## 4.1 Nonresponse Modelling

There are several challenges facing the CCHS that will be investigated over the next several months. The first issue is applying the nonresponse adjustments to the small samples collected over short time periods. One of the goals with the redesigned methodology is to produce a robust nonresponse model, but at the same time be as unbiased as possible. With the shorter time periods, it is unlikely that a robust model can be determined because of the limited sample size. In order to have a more robust model, the possibility of modelling the response mechanism using past years of CCHS data and applying it to the current data will be evaluated. This method would increase the efficiency of the weighting process since a new model would not need to be created for every file that is produced. Thus, weighting files could be produced in a timelier manner. As well, with the model being based on a large amount of data collected over one year, it is more stable and robust than a model based on a smaller amount of data collected in a short period of time. Under this proposed method, the nonresponse model for 6-month files would be based on larger samples. The nonresponse adjustments may still require some collapsing of groups to ensure that there are enough observations in an adjustment class and that large adjustments are avoided. This method has the disadvantage that it requires the assumption that the nonresponse mechanism remains constant over time.

## 4.2 User Expectations

With this idea of continuous collection, it must be made clear to users that the different weights and different files produced represent different time periods for the evolving population of interest. Differences between estimates could be the result of changes in the characteristics of the population or changes in the demographic composition of the population. By cumulating samples, users will get better estimates than those obtained from one sample alone in terms of variance. However, the cumulated estimates will not be for the same time period and the researcher will actually be estimating something different than that estimated from one sample. For example, by combining units collected in 2007 with those collected in 2008, an estimate with less variance can be calculated for the 2007-2008 collection period but this estimate will not correspond to the estimate for either the 2007 or 2008 collection periods. If the interest is to improve estimates for a particular time period then time series and small area estimation techniques may be useful but these estimation techniques can not be applied directly in the weighting steps.

## 4.3 Geography Changes

One of the challenges with continuous collection will be dealing with the ongoing changes in the geographical boundary definitions of health regions. Changes in boundaries are decided by provincial jurisdictions. The Health Statistics Division (HSD) has limited the changes to only once a year but sometimes these changes can be drastic. Along with this, some provinces use

different boundaries for the same time period. For example, Ontario is interested in both District Health Units and Local Health Integrated Networks. The disadvantage of these different geographies is that in order to integrate the weights from different time periods, a common geography will have to be defined and most likely will be the most recent one available. This common geography could simply be a variable added to the previous data files or weights could be recalibrated to updated projection counts for the new geographies for the previous time periods. At this point, the CCHS is shying away from adjusting the weights that have been previously released in favour of simply adding updated variables to the previous datasets. The weights will have been in the public domain for a long period of time and several analyses will have been published based on those weights. To revise the weights would only add doubt and confusion for the researchers using the data.

### 4.4 Variance Calculation

Variances for the CCHS estimates are estimated using the bootstrap method. In general, the same process that is applied to the sampling weights will be applied to the bootstrap weights. Bootstrap weights will have to be coordinated between the different time periods to ensure that the dependence between samples is properly reflected in the variance.

## 5 Conclusion

Although there are many advantages of the CCHS redesign, it creates several challenges in the weighting process. The idea of continuous collection and the proper weighting for such a design is a new challenge for the CCHS and the data users. Many of these issues were present with the original CCHS, but are more pronounced with the redesign. Given that the CCHS was generally conducted over a one year period, the changes over time were minor and somewhat ignored. This can not continue under the redesign where prolonged collection periods create a more pronounced time effect on the data collected. As well, the weighting process will have to be more efficient given the number of products being released. Changes to the weighting process itself and the idea of integrating existing weights to create weights for longer time periods should help with the timeliness of releasing the products. The proposed methodology will be in place for weighting of the 2007 CCHS and weights for the 2007 and 2008 releases will be integrated for the 2007-2008 time period.

## 6 References:

Alexander, Charles H. (2002). “Still Rolling: Leslie Kish’s “Rolling Samples” and the American Community Survey”. *Survey Methodology*, 28, 35-41.

Béland, Y., Dale, V. and Hamel, M. (2005). “Redesign of the Canadian Community Health Survey Program”. *Technical Report presented to Statistics Canada's Advisory Committee on Statistical Methods*, Meeting No. 41.

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*.

Chu, A., Brick, J.M., and Kalton, G.(1999). “Weights for Combining Surveys across Time or Space”. *Proceedings from the 1999 International Statistical Institute*, 103-104.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J.(1998). "Methodology of the Canadian Labour Force Survey." *Statistics Canada*.

Korn, E. L. and Graubard, B. I. (1995). *Analysis of Health Surveys.* Wiley.

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

Sarafin, C., Thomas, S., and Simard, M. (2007). "A Review of the Weighting Strategy for the Redesigned Canadian Community Health Survey". *To be published in the Proceedings of The Statistical Society of Canada 2007.*

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

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

Thomas, S. (2006). "Combining Cycles of the Canadian Community Health Survey". *To be published in the Proceedings of Statistics Canada Symposium 2006.*