CANADIAN COMMUNITY HEALTH SURVEY (CCHS) CYCLE 3.1 (2005)

PUBLIC USE MICRODATA FILE (PUMF) USER GUIDE

June 2006

Statistics Canada Statistique Canada

Canada

TABLE OF CONTENTS

1. Introduction

The Canadian Community Health Survey (CCHS) is a cross-sectional survey that collects information related to health status, health care utilization and health determinants for the Canadian population. The CCHS operates on a two-year collection cycle. The first year of the survey cycle ".1" is a large sample, general population health survey, designed to provide reliable estimates at the health region level. The second year of the survey cycle ".2" has a smaller sample and is designed to provide provincial level results on specific focused health topics.

This Public Use Microdata File (PUMF) contains data collected for CCHS Cycle 3.1 between January 2005 and December 2005. The CCHS Cycle 3.1 collects responses from persons aged 12 or older, living in private occupied dwellings in 122 health regions covering all provinces and territories. Excluded from the sampling frame are individuals living on Indian Reserves and on Crown Lands, institutional residents, full-time members of the Canadian Forces, and residents of certain remote regions. The CCHS covers approximately 98% of the Canadian population aged 12 and over.

This document has been produced to facilitate the manipulation of the CCHS Cycle 3.1 PUMF, which is described in detail in the following text and appendices.

Any questions about the data sets or their use should be directed to:

Electronic Products Help Line: 1 (800) 949-9491

For custom tabulations or general data support:
Client Custom Services, Health Statistics Division: (613) 951-1746
E-mail: hd-ds@statcan.ca

For remote access support: (613) 951-1653
E-mail: cchs-escc@statcan.ca
Fax: (613) 951-4198

2. Background

In 1991, the National Task Force on Health Information cited a number of issues and problems with the health information system. These problems were that: data was fragmented; data was incomplete; data could not be easily shared; data was not being analysed to the fullest extent; and the results of research were not consistently reaching Canadians[1].

In responding to these needs, the Canadian Institute for Health Information (CIHI), Statistics Canada and Health Canada joined forces to create a Health Information Roadmap.

From this mandate, the Canadian Community Health Survey (CCHS) was conceived. The format, content and objectives of the CCHS evolved through extensive consultation with key experts, federal, provincial and community health region stakeholders to determine their data requirements[2].

The purpose of this publication, the PUMF, is to follow through on the mandate of collecting reliable, relevant information on health services, health status, and health issues important to Canadians - at the regional, provincial and national level - and disseminating this information to the public.

[1] 1999. Health Information Roadmap Responding to Needs, Health Canada, Statistics Canada. p.3.
[2] 1999. Roadmap Initiative … Launching the Process. Canadian Institute for Health Information / Statistics Canada. ISBN 1-895581-70-2. p.19.

3. Objectives

The primary objectives of the .1 cycles of CCHS are to:

- Provide timely, reliable, cross-sectional estimates of health determinants, health status and health system utilization across Canada;

- Gather data at the sub-provincial levels of geography;

As a key component of the Population Health Survey Program of Statistics Canada, the .1 cycles of CCHS help fulfil broader requirements of health issues in Canada. These are:

- Aid in the development of public policy.

- Provide data for analytic studies that will assist in understanding the determinants of health.

- Collect data on the economic, social, demographic, occupational and environmental correlates of health.

- Increase the understanding of the relationship between health status and health care utilization.

4. Survey content

This section provides a general discussion of the consultation process used in survey content development and gives a summary of the final content selected for inclusion in CCHS Cycle 3.1. The second sub-section describes the common content in detail followed by a sub-section explaining the optional content of the CCHS Cycle 3.1.

4.1 Consultation processes

One of the main objectives of CCHS is to fill data gaps – in the areas of health determinants, health status and health system utilization - at the health region level.

To identify these gaps, a consultation process was conducted in Fall 2001 with more than 200 representatives of regional, provincial and federal government agencies as well as with the population health research community.

Consultations prior to CCHS Cycle 1.1 used a combination of qualitative and quantitative methods to identify the relative priority of broad topic areas. Prior to Cycle 2.1, the primary objective of the consultations was to identify new and emerging issues for which a data gap existed. For Cycle 3.1, the focus of consultations was to identify marginal improvements in existing questionnaire modules.

Based on these consultations, a list of topics to be included in Cycle 3.1 was drafted by Statistics Canada and approved by an Advisory Committee consisting of representatives from health regions, all provincial and territories ministries of health and Health Canada.

The final CCHS Cycle 3.1 questionnaire consisted of: approximately 25 minutes of common content, which was asked of all respondents; approximately 5 minutes of sub-sample content, in which some questionnaire modules were asked only of enough respondents to yield reliable estimates at the national and provincial level; and approximately 10 minutes of optional content.

Each health region was allocated 10 minutes of optional content. Regional representatives chose questionnaire modules from a fixed list according to local needs and priorities. Each optional content module was asked only of respondents living in the health regions who had selected the module.

4.2 Common content

Topics that make up the common content are varied, ranging from Alcohol, Exposure to Second-hand Smoke, through Physical Activities and Two-week Disability. Table 4.1 outlines the common content for the CCHS Cycle 3.1. These common content topics were asked of all respondents in all health regions.

Table 4.1 CCHS Cycle 3.1 common content modules

• Alcohol • Maternal experiences • Chronic conditions • Exposure to second-hand smoke • Flu shots • General health • Health care utilization • Height and weight • Injuries • Mammography • Sexual behaviour	• PAP smear test • Physical activities • Restriction of activities • Smoking • Two-week disability • Youth smoking • Income • Socio-demographic characteristics • Administration • Home care • Labour Force (short form)

4.3 Optional content

Some questionnaire modules were designated as optional so that regions could select modules related to their particular needs and priorities (see Table 4.2). It should be noted that, unlike the modules included in the common content, the resulting data from the optional content modules is not easily generalized across Canada.

4.4 Sub-sample content

Three sets of questionnaire modules were asked only of a subset of respondents. The aim of these modules was to permit calculation of provincial and national estimates while minimising response burden. In most cases, sub-samples were asked to three different groups of respondents. Content included in each of the sub-samples is listed in table 4.3.

See Appendix A for guidelines on using and interpreting sub-sample content.

Table 4.2 CCHS Cycle 3.1 optional topic modules

• Access to health care services* • Alcohol dependence ** • Blood pressure check • Breast examinations • Breast self examinations • Changes made to improve health • Childhood and adult stressors • Colorectal cancer exams • Contacts with mental health professionals • Dental visits • Depression • Diabetes care • Dietary supplement use ** • Distress • Driving and safety ** • Eating troubles ** • Eye examinations • Food choices • Food security • Fruit and vegetable consumption * • Health care system satisfaction • Health utility index * • Home safety ** • Hormone replacement therapy ** • Illicit drugs • Insurance coverage • Labour force – long form * • Leisure activities ** • Mastery ** • Medication use	• Mood ** • Smoking – stages of change • Nicotine dependence • Smoking cessation aids • Smoking – physician counselling • Ongoing problems ** • Oral health 1 ** • Oral health 2 • Patient satisfaction * • Physical check-up • Canadian Problem Gambling Index • Prostate cancer screening • Psychological well-being manifestation scale ** • Recent life events ** • Satisfaction with life • Sedentary activities • Self-esteem • Health status – SF-36 • Sense of coherence ** • Sleep • Social support – availability • Social support - utilization • Spiritual values** • Stress – coping • Stress – sources • Suicidal thoughts and attempts • Sun safety • Tobacco alternatives ** • Use of protective equipment • Voluntary organizations • Waiting times * • Work stress

* Sub-sample content, also available for selection by health regions as optional content.

** Available for optional content but not chosen.

Table 4.3 CCHS Cycle 3.1 Sub-sample modules

Sub-Sample 1 • Health Utility Index (HUI) • Fruit and Vegetable Consumption • Labour Force (long form). **Sub-Sample 2** • Measured Height and Weight	**Sub-Sample 3** • Access to Health Care Services • Waiting Times • Patient Satisfaction

5. Sample design

5.1 Target population

The CCHS targets persons aged 12 years and older who are living in private dwellings in the ten provinces and three territories. Persons living on Indian Reserves or Crown lands, those residing in institutions, full-time members of the Canadian Forces and residents of certain remote regions are excluded from this survey. The CCHS covered approximately 98% of the Canadian population aged 12 and older.

5.2 Health regions

For administrative purposes, each province is divided into health regions (HR) and each territory is designated as a single HR (Table 5.1). Statistics Canada, in consultation with the provinces, made minor changes to the boundaries of some of the HRs to correspond to the geography of the 2001 Census. During Cycle 3.1 of the CCHS, data was collected in 122 HRs in the ten provinces, in addition to one HR per territory, totalling 125 HRs.

Table 5.1 Number of health regions and targeted sample sizes by province/territory

Province	Number of HRs	Total sample size (targeted)
Newfoundland and Labrador	6	4,010
Prince Edward Island	4	2,000
Nova Scotia	6	5,040
New Brunswick	7	5,150
Quebec	16	24,280
Ontario	37	42,260
Manitoba	10	7,500
Saskatchewan	11	7,720
Alberta	9	12,200
British Columbia	16	16,090
Yukon	1	850
Northwest Territories	1	900
Nunavut	1	700
Canada	**125**	**128,700**

5.3 Sample size and allocation

To provide reliable estimates for the 125 HRs, and given the budget allocated to the CCHS Cycle 3.1 component, a sample of 128,700 respondents was desired. Although producing reliable estimates at the HR level was a primary objective, the quality of the estimates for certain key characteristics at the provincial level was also deemed important. Therefore, the sample allocation strategy, consisting of three steps, gave relatively equal importance to the HRs and the provinces. In the first two steps, the sample was allocated among the provinces according to their respective populations and the number of HRs they contained (Table 5.1). In the third step, each province's sample was allocated among its HRs proportionally to the square root of the estimated population in each HR.

This three-step approach guaranteed a sufficient sample for each HR with minimal disturbance to the provincial allocation of sample sizes. The sample sizes were increased before data collection to take into account out-of-scope and vacant dwellings and anticipated nonresponse. For the complete list of HRs and achieved sample sizes, see Section 9 on data quality.

Note that the three territories were not part of the above allocation strategy as they were dealt with separately. In total, 850 sample units were allocated to the Yukon, 900 to the Northwest Territories and 700 to Nunavut.

5.4 Frames, household sampling strategies

Cycle 3.1 of the CCHS used three sampling frames to select the sample of households: 49% of the sample of households came from an area frame, 50% came from a list frame of telephone numbers and the remaining 1% came from a Random Digit Dialling (RDD) sampling frame.

5.4.1 Sampling of households from the area frame

The CCHS used the area frame designed for the Canadian Labour Force Survey (LFS) as a sampling frame. The sampling plan of the LFS is a multistage stratified cluster design in which the dwelling is the final sampling unit[3]. In the first stage, homogeneous strata are formed and independent samples of clusters are drawn from each stratum. In the second stage, dwelling lists are prepared for each cluster and dwellings, or households, are selected from the lists.

For the purpose of the LFS 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 regrouped to create clusters. Some urban centres have separate strata for apartments or for census Enumeration Areas (EA) to pinpoint households with high income, immigrants and the native people. In each stratum, six clusters or residential buildings (sometimes 12 or 18 apartments) are chosen by a random sampling method with a probability proportional to size (PPS), the size of which corresponds to

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

the number of households. The number six is used throughout the sample design to allow a one-sixth rotation of the sample every month for the LFS.

The other cities and rural regions of each province are stratified first on a geographical basis, then according to socio-economic characteristics. In the majority of strata, six clusters (usually census EAs) are selected using the PPS method. Some geographically isolated urban centres are covered by a three-stage sampling design. This type of sampling plan is used for Quebec, Ontario, Alberta and British Columbia.

Once the new clusters are listed, the sample is obtained using a systematic sampling of dwellings. Table 5.2 gives an overview of the types of PSUs used for the entire LFS sample. 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. The yield of the sample, for example, is sometimes excessive. This particularly happens in sectors, for example, where there is an increase in the number of dwellings due to new construction. To reduce the cost of collection, an excessive output is corrected by eliminating, from the beginning, a part of the units selected and by modifying the weight of the sample design. Such a procedure, usually conducted at an aggregate level, is called *sample stabilization*. Moreover, the required sample size of households is increased to account for vacant dwellings, with experience having shown that 12% of all dwellings are not occupied by households that are part of the scope of observation (certain dwellings are vacant or occupied seasonally, others are occupied by households that are not targeted by the survey).

Table 5.2 Major first-stage units, sizes and yields

Area	Primary Sampling Unit (PSU)	Size (households per PSU)	Yield (sampled households)
Toronto, Montréal, Vancouver	Cluster	150-250	6
Other cities	Cluster	150-250	8
Most rural areas / small urban centres	Cluster	100-250	10

Requirements specific to the CCHS led to some modifications to this sampling strategy[4]. To get a base sample of 62 000 households for the CCHS, 86,000 dwellings must be selected from the area frame (to account for vacant dwellings and non-responding households). On an on-going monthly basis the LFS design provides approximately 60,000 dwellings distributed across the various economic regions in Canada whereas the CCHS requires a total of 86,000 dwellings distributed in the HRs, which have different geographic boundaries than those of the LFS economic regions. Overall, the CCHS required 43% more dwellings than those generated by the LFS selection mechanism, for an *adjustment factor* of 1.43 (86,000/60,000). However, at the HR level, the adjustment factors varied from 0.6 to 6.0.

[4] Morano, M., Lessard, S. and Béland, Y. (2000). Creation of a dual frame for the Canadian Community Health Survey, *2000 Proceedings of the Survey Methods Section*, Ottawa: Statistical Society of Canada, 249-254.

The changes made to the selection mechanism in a HR varied depending on the size of the adjustment factors. For HRs that had a factor smaller than or equal to 1, a simple stabilization, as described above, was 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 was repeated for all selected PSUs that were 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 for dwellings in a PSU, was repeated. For HRs with a factor between 4 and 6, the PSU sampling process was repeated not once but twice while that for dwellings was repeated only once. Where the chosen approach created an unnecessary surplus of dwellings, stabilization was performed.

It should be noted that the changes made to the LFS mechanism resulted in, at most, tripling the number of PSUs selected and, at most, doubling the number of dwellings selected in the PSUs, which explained the maximum adjustment factor of 6.0. At the HR level, adjustment factors were purposely capped at 6.0 for two reasons: to limit the listing of clusters (each newly 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 adjustment factor of certain HRs has consequently dictated the number of households required from the telephone frames.

Sampling of households from the area frame in the three territories

For operational reasons the area frame sample design implemented in the three northern territories had one additional stage of selection. For each territory, in-scope communities were first stratified based on various characteristics (population, geography, proportion of Inuit and/or Aboriginal persons, and median household income). There were five design strata in the Yukon, ten in the Northwest Territories and six in Nunavut. Then the first stage of selection consisted of randomly selecting one community with a probability proportional to population size within each design stratum. From that point on, the household sampling strategy from the area frame within the selected community was identical to the one described above.

It is worth mentioning that the frame for the CCHS covered 90% of the private households in the Yukon, 97% in the Northwest Territories and 71% in Nunavut.

5.4.2 Sampling of households from the list frame of telephone numbers

The list frame of telephone numbers was used in all but 5 HRs (the two RDD only HRs and the three territories) to complement the area frame. 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 administrative conversion files to obtain postal codes, and these were mapped to HRs to create list frame strata. There was one list frame stratum per HR. Within each stratum the required number of telephone numbers was selected using a simple random sampling process from the list. As for the RDD frame, additional telephone numbers were selected to account for the numbers not in service or out-of-scope. The hit rate observed under the list frame approach varied from 75% to 88% depending on the province, which was much higher that that for the RDD frame.

It is important to mention that the coverage of the list frame is less than the one for the RDD as unlisted numbers do not have a chance of being selected. Nevertheless, as the list frame is always used as a complement to the area frame, the impact of the undercoverage of the list frame is minimal and is dealt with in weighting.

5.4.3 Sampling of households from the RDD frame of telephone numbers

In four HRs, a Random Digit Dialling (RDD) sampling frame of telephone numbers was used to select the sample of households. The sampling of households from the RDD frame used the Elimination of Non-Working Banks (ENWB) method, a procedure adopted by the General Social Survey[5]. 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. The Canada Phone Directory and telephone companies' billing address files were used in conjunction with various internal administrative files to eliminate non-working banks.

Using available geographic information (postal codes), the banks on the frame were regrouped to create RDD strata to encompass, as closely as possible, the HR areas. Within each RDD stratum, a bank was randomly chosen and a number between 00 and 99 was generated at random to create a complete, ten-digit telephone number. This procedure was repeated until the required number of telephone numbers within the RDD stratum was reached. Frequently, the number generated is not in service or is 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. Within the CCHS, the hit rates ranged from 27% to 49% among the four HRs which required the use of the RDD frame.

5.5 Sampling of interviewees

As was done for the previous cycles, the selection of individual respondents was designed to ensure over-representation of youths (12 to 19). The selection strategy was designed to consider user needs, cost, design efficiency, response burden and operational constraints. For the CCHS Cycle 3.1, it was decided to select one person per household using varying probabilities taking into account the age and the household composition. Many scenarios based on various parameters were simulated in order to determine the optimal approach without causing extreme sampling weights.

Table 5.3 gives the selection weight multiplicative factors used to determine the probabilities of selection of individuals in sampled households by age. As an example, for a three-person household (two 45-64 adults and one 15-year-old), the teenager would have 5 times the chance of being selected compared to the adults. To avoid extreme sampling weights, there is one exception

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

to this rule: if the size of the household is greater than or equal to 5 or the number of 12-19 year olds is greater than or equal to 3 then the selection weight multiplicative factor equals 1 for each individual in the household. Consequently, all people in that household have the same selection probability.

Table 5.3 Selection weight multiplicative factor for person-level sampling strategy by age

	Selection Weight Multiplicative Factor				
Age	**12-19**	**20-29**	**30-44**	**45-64**	**65+**
Factor	5	2	2	1	1

5.6 Sample allocation over the collection period

In order to balance interviewer workload and to minimize possible seasonal effects on estimates of certain key characteristics such as physical activity, the initial sample of dwellings / telephone numbers was allocated at random, within each HR, over the 11 months of data collection (the 12th month was a clean-up month). To start with, each PSU selected in the first stage from the area frame was randomly assigned to a collection quarter (Q1: January to March 2005, Q2: April and May 2005, Q3: June to August 2005 and Q4: September to November 2005). Within each collection quarter, the selected dwellings were randomly allocated to a collection month. For the telephone frames, independent samples were selected each month. This strategy ensured that each CCHS quarterly sample was representative of the Canadian in-scope population.

5.7 Supplementary buy-in sample in three health regions in Quebec

In order to allow for more reliable estimates for sub-regional areas, three health regions in the province of Quebec provided extra funds so that a larger sample of dwellings could be selected. The buy-in sample was combined with the main sample to produce one large data file.

The entire buy-in sample was selected from the list frame of telephone numbers. The Canada Phone Directory was linked to internal administrative files in order to stratify the listed telephone numbers in sub-regional areas (8 for Bas-St-Laurent, 12 for Montréal-Centre and 2 for Laval). The sample size per sub-regional area was based upon the funding available and the requirements of the health region to obtain reliable estimates by sub-regional area (Bas St-Laurent added 2,400 sample units, Montréal-Centre added 2,295 and Laval added 1,080). Table 5.4 gives the sample allocation by sub-regional area.

Table 5.4 Extra unit allocation in Bas-St-Laurent, Montréal-Centre and Laval health regions

Sub-regional area	Total sample size (targeted)
Région de Bas-St-Laurent	**2,400**
La Matapédia	300
Matane	300
La Mitis	300
Rimouski-Neigette	300
Les Basques	300
Rivière-du-Loup	300
Témiscouata	300
Kamouraska	300
Région de Montréal-Centre	**2,295**
Pierrefonds et Lac St-Louis	191
LaSalle et du Vieux Lachine	191
Verdun/Côte-St-Paul, St-Henri et Pointe-St-Charles	191
René-Cassin, NDG/Montréal-Ouest	191
Côte-des-Neige, Métro et Parc Extension	191
Nord de l'île et St-Laurent	191
Ahuntsic et Montréal-Nord	191
Petite-Patrie et Villeray	191
Des Faubourgs, Plateau Mont-Royal et St-Louis-du Parc	191
St-Michel et St-Léonard	191
Hochelaga-Maisonneuve, Olivier-Guimond et Rosemont	191
Rivière-des-Prairies, Mercier-Est/Anjou et Pointe-aux-Trembles/Montréal-Est	191
Région de Laval	**1,080**
Est	540
Ouest	540

6. Data collection

6.1 Computer-assisted interviewing

Collection for CCHS Cycle 3.1 took place from January to December 2005. Over the collection period, a total of 132,947 valid interviews were conducted using computer assisted interviewing (CAI). Approximately half the interviews were conducted in person using computer assisted personal interviewing (CAPI) and the other half were conducted over the phone using computer assisted telephone interviewing (CATI).

CAI offers two main advantages over other collection methods. First, CAI offers a case management system and data transmission functionality. This case management system automatically records important management information for each attempt on a case and provides reports for the management of the collection process. CAI also provides an automated call scheduler, i.e. a central system to optimise the timing of call-backs and the scheduling of appointments used to support CATI collection.

The case management system routes the questionnaire applications and sample files from Statistics Canada's main office to regional collection offices (in the case of CATI) and from the regional offices to the interviewers laptops (for CAPI). Data returning to the main office takes the reverse route. To ensure confidentiality, the data is encrypted before transmission. The data are then unencrypted when they are on a separate secure computer with no remote access.

Second, CAI allows for custom interviews for every respondent based on their individual characteristics and survey responses. This includes:

- questions that are not applicable to the respondent are skipped automatically
- edits to check for inconsistent answers or out-of-range responses are applied automatically and on-screen prompts are shown when an invalid entry is recorded. Immediate feedback is given to the respondent and the interviewer is able to correct any inconsistencies.
- question text, including reference periods and pronouns, is customised automatically based on factors such as the age and sex of the respondent, the date of the interview and answers to previous questions.

6.2 CCHS application development

For Cycle 3.1, two separate CAI applications were developed for telephone interviews (CATI) and personal interviews (CAPI). This was done in order to take advantage of newly standardized entry and exit application components/procedures developed at Statistics Canada, and to customise each applications' functionality to the type of interview being conducted. The applications consisted of entry, health content (known as the C2), and exit components.

Entry and exit components contain standard sets of questions designed to guide the interviewer through contact initiation, collection of important sample information, respondent selection and determination of cases status. The C2 consists of the health modules themselves and made up the

bulk of the applications. This includes common modules asked of all respondents and optional content which differed by health region.

A subset of the C2 modules, along with the entry and exit components was then used to create two test applications. Each application was pilot tested separately during the late summer of 2004.

The main objectives of the pilot tests were:

- To test changes made to the entry, exit and C2 components
- To evaluate respondent reaction to the new questions introduced in cycle 3.1 on Diabetes and Hormone Replacement Therapy
- To test the technical infrastructure, including transmission of data on the servers, of the regional offices and procedures unique to CATI BLAISE interviewing

Feedback from the pilot tests was used to make modification/improvements to the pilot applications. Once these modifications were complete, final testing of the full applications began. This consisted of three stages of internal testing: block, integrated and end to end.

Block level testing consists of independently testing each content module or "block" to ensure skip patterns, logic flows and text, in both official languages, are specified correctly. Skip patterns or logic flows across modules are not tested at this stage as each module is treated as a stand alone questionnaire. Once all blocks are verified by several testers they are added together along with entry and exit components into integrated applications. These newly integrated applications are then ready for the next stage of testing.

Integrated testing occurs when all of the tested modules are added together, along with the entry and exit components, into an integrated application. This second stage of testing ensures that key information such as age and gender are passed from the entry to the C2 and exit components of the applications. It also ensures that variables affecting skip patterns and logic flows are correctly passed between modules within the C2. Since, at this stage the applications essentially function as they will in the field, all possible scenarios faced by interviewers are simulated to ensure proper functionality. These scenarios test various aspects of the entry and exit components including, establishing contact, collecting contact information, determining whether a case is in scope, rostering households, creating appointments and selecting respondents. The applications are also tested to ensure that during an interview, correct modules are triggered reflecting health region optional content selections.

End to end testing occurs when the fully integrated applications are placed in simulated collection environment. The applications are loaded onto computers that are connected to a test server. Data is then collected, transmitted and extracted in real time, exactly as it would be done in the field. This last stage of testing allows for the testing of all technical aspects of data input, transmission and extraction for each of the CCHS 3.1 applications. It also provided a final chance of finding errors within the entry, C2 and exit components.

6.3 Interviewer training

Project managers from regional collection offices attended CCHS cycle 3.1 training sessions at Statistics Canada during December 2004. These sessions were conducted by the CCHS project team and were used to outline the interviewer training courses to be used in the regions. Project mangers then returned to the regions and conducted customised training sessions for their interviewing staff. Members of the survey team attended these training sessions to offer additional support and clarify any questions or concerns that may have arisen.

The focus of these sessions were to get interviewers comfortable using the CCHS 3.1 applications and familiarise interviewers with survey content. The training sessions focused on:

- goals and objectives of the survey
- survey methodology
- application functionality
- review of the questionnaire content and exercises
- interviewer techniques for maintaining response - complete exercises to minimise non-response
- use of mock interviews to simulate difficult situations and practise potential non-response situations
- survey management
- transmission procedures

One of the key aspects of the training was a focus on minimizing non-response. Exercises to minimise non-response were prepared for interviewers. The purpose of these exercises was to have the interviewers practice convincing reluctant respondents to participate in the survey. There was also a series of refusal avoidance workshops given to the senior interviewers responsible for refusal conversion in each regional collection office.

6.4 The interview

Sample units selected from the telephone list and RDD frames were interviewed from centralised call centres using CATI. The CATI interviewers were supervised by a senior interviewer located in the same call centre. Units selected from the area frame were interviewed by decentralised field interviewers using CAPI. While in some situations field interviewers were permitted to complete some or part of an interview by telephone, almost three-quarters (72.9%) of these interviews were conducted exclusively in person. CAPI interviewers worked independently from their homes using laptop computers and were supervised from a distance by senior interviewers. The variable SAME_TYP on the PUMF indicates whether a case was selected from the area frame (CAPI) or from the telephone or RDD frame (CATI).

In all selected dwellings, a knowledgeable household member was asked to supply basic demographic information on all residents of the dwelling. One member of the household was then selected for a more in-depth interview, which is referred to as the C2 Interview.

CAPI interviewers were trained to make an initial personal contact with each sampled dwelling. In cases where this initial visit resulted in non-response, telephone follow-ups were permitted. The variable ADME_N09 indicates whether the interview was completed face-to-face, by telephone or using a combination of the two techniques.

To ensure the quality of the data collected, interviewers were instructed to make every effort to conduct the interview with the selected respondent in privacy. In situations where this was unavoidable, the respondent was interviewed with another person present. Flags on the PUMF indicate whether somebody other than the respondent was present during the interview (ADME_N10) and whether the interviewer felt that the respondent's answers were influenced by the presence of the other person (ADME_N11).

To ensure the best possible response rate attainable, many practices were used to minimise non-response, including:

Introductory letters

Before the start of each collection period introductory letters explaining the purpose of the survey were sent to the sampled households. These explained the importance of the survey and provided examples of how CCHS Cycle 3.1 data would be used.

Initiating contact

Interviewers were instructed to make all reasonable attempts to obtain interviews. When the timing of the interviewer's call (or visit) was inconvenient, an appointment was made to call back at a more convenient time. If requests for appointments were unsuccessful over the telephone, interviewers were instructed to follow-up with a personal visit. If no one was home on first visit, a brochure with information about the survey and intention to make contact was left at the door. Numerous call-backs were made at different times on different days.

Refusal conversion

For individuals who at first refused to participate in the survey, a letter was sent from the nearest Statistics Canada Regional Office to the respondent, stressing the importance of the survey and the household's collaboration. This was followed by a second call (or visit) from a senior interviewer, a project supervisor or another interviewer to try to convince respondent of the importance of participating in the survey.

Language barriers

To remove language as a barrier to conducting interviews, each of the Statistics Canada Regional Offices recruited interviewers with a wide range of language competencies. When necessary, cases were transferred to an interviewer with the language competency needed to complete an interview. In addition, the survey questions were translated into the following languages: Chinese, Punjabi and Inuktitut. Chinese and Punjabi were the most common language barriers identified by the regional offices. The Inuktitut translation was used to facilitate collection in Nunavut.

Youth interviews

Interviewers were obliged to obtain verbal permission from parents/guardians to interview youths between the ages of 12 to 15 who were selected for interviews. Several procedures were followed by interviewers to alleviate potential parental concerns and to ensure a completed interview. Interviewers carried with them a card entitled "Note to parents / guardians about interviewing youths for the Canadian Community Health Survey". This card explained the purpose of collecting information from youth, lists the subjects to be covered in the survey, asks for permission to share and link the obtained information and explains the need to respect a child's right to privacy and confidentiality.

If a parent/guardian asked to see the actual questions; interviewers were instructed to either show the survey questions, or if the interviewer was being conducted by phone, to immediately have the regional office send a copy of the questionnaire.

If privacy could not be obtained to interview the selected youth either in person or over the phone (another person listening in) the interview was coded a refusal. However, for CAPI interviews, if privacy could not be obtained to interview the selected youth, the interviewer was able to propose to the parent/guardian that the interviewer read the questions out loud and the youth enter their answers directly on the computer.

During all interviews conducted with youths, survey questions regarding income and food security were answered by the parent/guardian. These questions were asked at the end of the survey questionnaire, so that when they came up, the parent/guardian could complete the interview.

Proxy interviews

In cases where the selected respondent was, for reasons of physical or mental health, incapable of completing an interview, another knowledgeable member of the household supplied information about the selected respondent. This is known as a proxy interview. While proxy interviewees were able to provide accurate answers to most of the survey questions, the more sensitive or personal questions were beyond the scope of knowledge of a proxy respondent. This resulted in some questions from the proxy interview being unanswered. Every effort was taken to keep proxy interviews to a minimum. The variable ADME_PRX indicates whether a case was completed by proxy.

6.5 Field operations

The majority of the cycle 3.1 sample was divided into eleven two-month overlapping collection periods. Regional collection offices were instructed to use the first 4 weeks of each collection period to resolve the majority of the sample, with next 4 weeks being used finalise the remaining sample and to follow up on outstanding non-response cases. All cases were to have been attempted by the second week of each collection period.

Certain portions of the 3.1 sample had required slightly different collection approaches. Separate *quarterly* collection periods were created to facilitate work in the remote communities in Nunavut, NWT and the Yukon. The collection of the measured height and weight sub-samples (see Appendix A) were also divided into quarterly collection periods to ensure that specialised collection techniques were used properly and consistently and ensure that specially trained interviewers were assigned to that sample. Finally, quarterly collection periods were used to manage collection of RDD sample as it had different characteristics from the list frame, including different collection targets and timeframes.

Sample files were sent approximately two weeks before the start of each collection period to centralised collection offices. A series of dummy cases were included with each sample. These cases were completed by senior interviewers for the purposes of ensuring that all data transmission procedures were working through the collection cycle. Once, the samples were received, project supervisors were responsible for planning CAPI interviewer assignments. Wherever possible, assignments were generally no larger than 15 cases per interviewer.

Transmission of cases from each of the CATI offices to head office was the responsibility of the regional office project supervisor, senior interviewer and the technical support team. These transmissions were performed nightly and sent all completed cases to Statistics Canada's head office Completed CAPI interviews were transmitted daily from the interviewer's home directly to Statistics Canada's head office using a secure telephone transmission.

In June and again in November, non-response cases (excluding refusals) from previous collection periods were returned for further collection activities. These consisted of situations where the selected respondents had been absent for the duration of the initial collection period. These non-response cases were again approached and encouraged to participate in the survey.

At the end of data collection, a national response rate of 79% was achieved. Complete details regarding the response rates can be found in Section 9.

6.6 Quality control and collection management

During the 3.1 collection cycle, several methods were used to ensure data quality and to optimize collection. These included using internal measures to verify interviewer performance and the use of a series of ongoing reports to monitor various collection targets and data quality.

A system of validation was used for CAPI cases whereby interviewers had their work validated on a regular basis by the Regional Office. Each collection period, randomly selected cases were flagged in the sample. Regional office managers and supervisors created lists of cases to be validated. These cases were handed to the validation team who then contacted households to verify that a legitimate interview took place. Validation procedures generally occurred during the first few weeks of a collection period to ensure that any issues were detected promptly. Interviewers were provided feedback by their supervisors on a regular basis.

CATI interviewers were also randomly chosen for validation. Validation in the CATI collection offices consisted of senior interviewers monitoring interviews to ensure proper techniques and procedures (reading the questions as worded in the applications, not prompting respondents for answers, etc.) were followed by the interviewer.

A series of reports were produced to effectively track and manage collection targets and to assist in identifying other collection issues.

Cumulative reports were generated at the end of each collection period, showing response, link, share and proxy rates for both the CATI and CAPI samples by individual health region. The reports were useful in identifying health regions that were below collection target levels, allowing the regional offices to focus efforts in these regions. In addition to these cumulative reports, a series of reports outlining weekly collection targets for each regional office to assist in resource planning and achieving their collection targets were used.

Using information obtained from the CAI applications, further analysis was done in head office in order to identify interviews that were completed below acceptable time frames. These short interviews were flagged, removed from the microdata and treated as non-response.

Customised reports were also created and used to examine specific data quality issues that arose during collection. For example, one of the key collection targets for the measured height and weight sub-sample was the rate at which interviewers were obtaining a valid height and weight measure. This was a very important rate that was not evident in the cumulative reports. Once the rates were found to be under target, custom reports were created and used to identify interviewers who were having troubles achieving valid measurements. These interviewers then received additional training reviewing special procedures related to the collection of measured height and weight data.

7. Data processing

7.1 Editing

Most editing of the data was performed at the time of the interview by the computer-assisted interviewing (CAI) application. It was not possible for interviewers to enter out-of-range values and flow errors were controlled through programmed skip patterns. For example, CAI ensured that questions that did not apply to the respondent were not asked.

In response to some types of inconsistent or unusual reporting, warning messages were invoked but no corrective action was taken at the time of the interview. Where appropriate, edits were instead developed to be performed after data collection at Head Office. Inconsistencies were usually corrected by setting one or both of the variables in question to "not stated".

7.2 Coding

Pre-coded answer categories were supplied for all suitable variables. Interviewers were trained to assign the respondent's answers to the appropriate category.

In the event that a respondent's answer could not be easily assigned to an existing category, several questions also allowed the interviewer to enter a long-answer text in the "Other-specify" category. All such questions were closely examined in head office processing. For some of these questions, write-in responses were coded into one of the existing listed categories if the write-in information duplicated a listed category. For all questions, the 'Other-specify' responses are taken into account when refining the answer categories for future cycles.

7.3 Creation of derived and grouped variables

To facilitate data analysis and to minimise the risk of error, a number of variables on the file have been derived using items found on the CCHS Cycle 3.1 questionnaire. Derived variables generally have a "D", "G" or "F" in the fifth character of the variable name. In some cases, the derived variables are straightforward, involving collapsing of response categories. In other cases, several variables have been combined to create a new variable. The *Derived Variables Documentation (DV)* provides details on how these more complex variables were derived. For more information on the naming convention, please go to Section 12.2.

7.4 Weighting

The principle behind estimation in a probability sample such as the CCHS Cycle 3.1 is that each person in the sample "represents", besides himself or herself, several other persons not in the sample. For example, in a simple random 2% sample of the population, each person in the sample represents 50 persons in the population. In the terminology used here, it can be said that each person has a weight of 50.

The weighting phase is a step that calculates, for each person, his or her associated sampling weight. This weight appears on the PUMF, and must be used to derive meaningful estimates from the survey. For example, if the number of individuals who smoke daily is to be estimated, it is done by selecting the records referring to those individuals in the sample having that characteristic and summing the weights entered on those records.

Details of the method used to calculate sampling weights are presented in Section 8.

7.5 Conversion of CCHS 3.1 Master File to Public Use Microdata File (PUMF)

The approach for creating a PUMF is intended to balance the requirements for maintaining respondent confidentiality by minimising disclosure risks, while providing the most useful data at the level of geography of the health region. The following paragraphs outline some of the procedures applied to convert the CCHS master file into a PUMF.

Health regions: Health regions fall under provincial jurisdiction. As mentioned in Section 5.2, at the time of the design of the 3.1 sample, there were 125 HRs across Canada. During collection, Newfoundland and Labrador rolled up two health regions with two others and Ontario split one and merged the two parts into two existing health regions, thereby leaving 122 HRs to be disseminated for the CCHS 3.1 release. Further, three HRs in Quebec bought extra sample units in order to obtain sub-regional estimates. These HRs are Montréal (12 CLSC sub-regions), Bas-Saint-Laurent (8) and Laval (2). Thus, the master data file provides representative information for 122 HRs and for 22 CLSC buy-ins.

Some health regions have population sizes that were considered to be too small to appear individually in the PUMF. These health regions were thus collapsed with other(s). The approach for keeping or collapsing health region geography in the PUMF entails applying a minimum value of 70,000 on the population size. This resulted in:

i) collapsing of 33 HRs, in all provinces except Quebec and Nova Scotia, into 15 health region groupings,
ii) grouping of the three territories into a single entity (results in some optional content suppression),
iii) one Quebec HR (Nord-du-Québec) is excluded from the PUMF because its small population size and demographic characteristics did not lend itself well to being collapsed with other HRs,
iv) all sample design variables were excluded.

After collapsing, the CCHS PUMF comprised 101 geographic HRs/HR groupings across the country. HRs before and after collapsing are shown in Appendix B. CLSCs/CLSC groupings are included to provide data in keeping with the additional sample in the three Quebec HRs. All CLSC's in Laval and Montréal have population counts over 70,000 and therefore did not require any collapsing. The 8 CLSC's in the Région Bas-St-Laurent were collapsed into two regions as indicated in Appendix B. Therefore the PUMF provides information on 101 HRs/HR groupings and for 16 CLSC/CLSC groupings.

Optional content suppression: As a result of the grouping of HRs, the optional content that was not common to each health region within a grouping would have to be suppressed. Although optional content selection could vary from region to region within a province, in Cycle 3.1 all health regions within each province selected the same optional content; therefore this suppression was not necessary for any provinces. However, because the three territories were grouped, suppression of optional content occurs in the territories if only one or two of the territories selected the content.

Disclosure control: As mentioned earlier, the 3.1 PUMF is designed to preserve analytical value of data as much as possible while ensuring the potential for identifying individual respondents is minimal.

Several controls were implemented in creating the PUMF. Restriction methods such as removing direct identifiers (e.g., sample ID, name, telephone number), reducing, recoding, and/or suppressing detail based on small frequencies given specific socio-demographic characteristics were used. Examples of some master file variables not being included in the PUMF due to a high risk of disclosure (either because the variable is a risk on its own or is so in combination with other variables) include: attempted suicide in the past 12 months, exact height and weight data and whether the respondent is presently pregnant.

Some response categories deemed to be at possible risk of disclosure were regrouped and included on the PUMF. Examples include: body part affected by most serious injury, number of times consulting health professionals, number of years since stopping smoking and main source of household income. Thus, the PUMF contains both fewer variables and less detail compared with the 3.1 master file, but in a way that preserves analytical value to the data.

Age of respondent: Respondents' age is provided in the CCHS 3.1 PUMF in age groupings. Most are found in 5-year age groups, from age 20-24 to age 75-79. All respondents aged 80 and over were also grouped. In previous CCHS x.1 PUMFs, youth age groups included those aged 12-14 and 15-19. In this current PUMF, the older youth age group has been split between those aged 15-17 and those aged 18-19. The older youth age groupings were changed for two principal reasons:

1) Body Mass Index (BMI) has emerged as a top priority in the public health realm. New BMI calculations are newly available for youths aged 12-17 but would not have been fully available to users via the older youth age grouping;

2) in prior releases we had to suppress data for those aged 18-19 for variables which were only asked to respondents aged 18 and over.

Therefore, the proposed new youth age groupings have at least three important benefits:

i) BMI data can be presented for those aged 12-17;

ii) data for those aged 18-19 is longer automatically suppressed for variables that were only asked to those aged 18 and over;

iii) users can still compare youth data from previous PUMFs by grouping the 15-17 year olds with the 18-19 year olds to create the age group 15-19.

Although information on some variables were collected for people of a specific age bracket, some data for certain ages still had to be suppressed to ensure confidentiality. For example, labour force data were collected for respondents aged 15 to 75. Because there is an age group 75-79 on the PUMF, publishing labour force data for 75-year olds would reveal their exact age by default. Therefore, labour force data (LBSE) are only available on the PUMF for respondents aged 15-74. Similar suppression was done for education (EDUE) data of 14-year olds and maternal experience (MEXE) data for 55-year old women.

Special suppression was done for maternal experience (MEXE) data for women aged 15-17 because there was concern for a high risk of disclosure for women in this age group. As such, though MEXE data on the master file are shown for all women aged 15 to 55, the PUMF only presents data for women aged 18 to 54.

8. Weighting

In order for estimates produced from survey data to be representative of the covered population, and not just the sample itself, users must incorporate the survey weights in their calculations. A survey weight is given to each person included in the final sample, that is, the sample of persons having answered the survey. This weight corresponds to the number of persons in the entire population that are represented by the respondent.

As described in Section 5, CCHS Cycle 3.1 had recourse to three sampling frames for its sample selection: an area frame acting as the primary frame and two frames formed of telephone numbers complementing the area frame. Since only minor differences differentiate the two frames formed of telephone numbers in terms of weighting, they are treated together. They are referred to as being part of the telephone frame.

The weighting strategy was developed by treating both the area and telephone frames independently. Weights resulting from these two frames are afterwards combined into a single set of weights through a step called "*integration*". After some adjustments, this integrated weight becomes the final weight. Note that depending on the need, one or two frames were used for the selection of the sample within a given health region (HR). The weighting strategy deals with this aspect at the integration step.

8.1 Sample weighting

As mentioned previously, units from both area and telephone frames are treated separately up to the integration step. Sub-section 8.1.1 provides details on the weighting strategy for the area frame, while sub-section 8.1.2 deals with the strategy for the telephone frame. The integration of the two frames is discussed in 8.1.3. This is followed by the two last weighting steps, that is, the adjustment controlling for the seasonal effect and the post-stratification, which are explained in sub-sections 8.1.4 and 8.1.5 respectively.

Although these two frames were used to cover the three territories, some modifications had to be done relative to their use. These modifications affected the weighting of these three regions substantially, and they are reported in sub-section 8.1.6.

Diagram A presents an overview of the different adjustments that are part of the weighting strategy, in the order in which they are applied. A numbering system is used to identify each adjustment applied to the weight and will be used throughout the section. Letters A and T are used as prefixes to refer to adjustments applied to the units on the Area and Telephone frames respectively, while prefix I identifies adjustments applied from the Integration step.

Diagram A Weighting strategy overview

Area frame	Telephone frame
A0 – Initial weight	T0 – Initial weight
A1 – Sample increase	T1 – Number of months
A2 – Stabilization	T2 – Removal of out-of-scope numbers
A3 – Removal of out-of-scope units	T3 – Coverage of the list frames
A4 – Household nonresponse	T4 – Combination of the list frames
A5 – Creation of person level weight	T5 – Household nonresponse
A6 – Person nonresponse	T6 – Households without telephone
Final area frame weight	T7 – Creation of person level weight
↘	T8 – Person nonresponse
	T9 – Multiple lines
	Final telephone frame weight
	↙
I1 – Integration	
I2 – Seasonal effect	
I3 - Post-stratification	
Final CCHS Cycle 3.1 weight	

8.1.1 Weighting of the area frame sample

A0 – Initial weight

Since the mechanism established for the LFS was used to select the area frame sample, the initial weights had to be computed with respect to that mechanism. First, within each stratum defined by the LFS, clusters (primary sampling units) are selected with probabilities proportional to the number of households (based on 2001 Census counts). Next, dwellings are sampled within each selected cluster using systematic sampling. The product of the probabilities for each of these selections represents the overall probability of selection, and the inverse of that probability is used as the CCHS Cycle 3.1 initial weight. For more details about the selection mechanism, as well as a more complete definition of strata and clusters, refer to Statistics Canada (1998)[6].

A1 – Sample increase

Some modifications were made to the default LFS mechanism at the time of sample selection for CCHS Cycle 3.1. The LFS design provides approximately 60,200 dwellings nationally, but CCHS Cycle 3.1 requirements in terms of sample size were higher for some regions. Modifications were made in order to obtain extra sample within those HRs requiring more sample. More specifically, these modifications consisted of repeating the sampling process within all selected clusters of the HR. This had the effect of boosting the sample size and had to be taken into consideration in the weighting by adjusting the probability of selection. An adjustment factor, A1, representing the

[6] Statistics Canada. 1998. *Methodology of the Canadian Labour Force Survey.* Statistics Canada. Cat. No. 71-526-XPB.

sample increase rate, was calculated. However, a sample increase was not required in every HR. In some regions, the LFS design provided more sample than needed by CCHS Cycle 3.1. For those regions, the adjustment factor represents a sample decrease instead of representing a sample increase. The initial weight, A0, is multiplied by this adjustment factor, resulting in weight A1.

A2 – Stabilization

In some HRs, increasing the sample as described in the previous paragraph, resulted in a significantly larger sample than necessary. Stabilization was therefore instituted to bring the sample size back down to the desired level. The stabilization process consisted of randomly sub-sampling dwellings at the HR level. An adjustment factor representing the effect of this stabilization was calculated in order to adjust the probability of selection appropriately. This factor, multiplied by weight A1, produces weight A2.

A3 – Removal of out-of-scope units

Among all dwellings sampled, a certain proportion is identified during collection as being out-of-scope. Dwellings that are demolished or under construction, vacant, seasonal or secondary, and institutions are examples of out-of-scope cases for CCHS Cycle 3.1. These dwellings were simply removed from the sample, leaving only a sample that consisted of in-scope dwellings. These dwellings maintain the same weight as in the previous step, which is now called A3.

A4 – Household nonresponse

During collection, a certain proportion of interviewed households inevitably resulted in nonresponse. This usually occurs when a household refuses to participate in the survey, provides unusable data, or cannot be reached for an interview. Weights of the non-responding households were distributed using response propensity classes to the responding households. The CHAID (Chi-Square Automatic Interaction Detector) algorithm, available in Knowledge Seeker[7], was used to identify the best characteristics to divide the sample into groups that were dissimilar with respect to response/nonresponse. Note that the groups were formed independently within each HR. Since the information available for non-respondents is limited, only characteristics such as collection period and a rural/urban indicator could be used for the creation of the classes. Analysis concluded that the rural/urban indicator was the most significant characteristic. The rural/urban indicator was also significant (with 4 periods: January to March, April to May, June to August, and September to November) for the creation of classes in each HR. An adjustment factor was calculated within each class as follows:

$$\frac{\textit{Sum of weight A3 for all households}}{\textit{Sum of weight A3 for all responding households}}$$

[7] ANGOSS Software. 1995. Knowledge Seeker IV for Windows - User's Guide. ANGOSS Software International Limited.

Weight A3 was multiplied by this factor to produce weight A4 for the responding households. Non-responding households were dropped from the process at this point.

A5 – Creation of person level weight

Since persons are the desired sampling units, the household level weights computed to this point need to be converted down to the person level. This weight is obtained by multiplying weight A4 by the inverse of the probability of selection of the person selected in the household. This gives the weight A5. As mentioned earlier, the probability of selection for an individual changes depending on the number of people in the household and the ages of those individuals (see Section 5.5 for more details).

A6 – Person nonresponse

A CCHS Cycle 3.1 interview can be seen as a two-part process. First the interviewer gets the complete roster of the people living within the responding household. Second, (s)he interviews the selected person within the household. In some cases, interviewers can only get through the first part, either because they cannot get in touch with the selected person, or because that selected person refuses to be interviewed. Such cases are defined as person nonresponse and an adjustment factor must be applied to the weights of respondents to account for this nonresponse. Using the same methodology as was used in the treatment of household nonresponse, the adjustment was applied within classes based on characteristics available for both respondents and non-respondents. All characteristics collected when creating the roster of household members were in fact available for the creation of the classes. The CHAID algorithm, available in Knowledge Seeker, was used to define the classes. Note that groups were formed independently within each HR. Depending on the HR, the following characteristics were used to form the adjustment classes: household size groups, urban/rural indicator, collection period, number of persons 12 years of age or older, living arrangement, sex, marital status, and age resulting in the following adjustment factor:

$$\frac{\textit{Sum of weight A5 for all selected persons}}{\textit{Sum of weight A5 for all responding selected persons}}$$

Weight A5 for responding persons was multiplied by the above adjustment factor to produce weight A6. Non-responding persons were dropped from the weighting process from this point onward.

Since this adjustment was the last one necessary for the sample drawn from the area frame, weight A6 represents the ***final area frame weight***. This weight is later integrated with the final weight of the telephone frame (section 8.1.3) to create the final CCHS Cycle 3.1 weight.

8.1.2 Weighting of the telephone frame sample

As mentioned earlier, the telephone frame is composed of two frames: a Random Digit Dialling (RDD) frame and a list frame, of which only one can be used in a specific HR. When the list frame

is used, it is always used as a complement to the area frame within the HR. When the RDD frame is used, it is always used as the only frame for the HR. However, for the purposes of weighting, units coming from these two frames are treated together and therefore are subject to the same adjustments. There are three exceptions: first, since the initial probability of selection of a dwelling is relative to the frame used for the selection, this probability will be slightly different depending on whether the unit is from the RDD frame or the list frame. The other exceptions concern the adjustments T3 and T4. Details about these exceptions are given in the sub-sections presenting the adjustments implicated.

There is another aspect particular to units coming from the telephone frame that affects the way the sample was weighted. This particularity concerns the geographical location of sampled units. The geographical boundaries used to select the sample from the telephone frame did not perfectly replicate the HR geography. Consequently, some units were selected from one location while the information collected at the time of the interview placed them in a neighbouring region. This particularity was handled in the weighting by applying all adjustments relative to the HR assigned at the time of sample selection. However, since it is required that all units belong to their correct HR, that is, the HR identified during collection, all unit weights were adjusted according to the correct HR based on information from the respondent. This adjustment was incorporated during post-stratification (I3), described later in this section.

T0 –Initial weight

The initial weight is computed differently between the RDD and list frame samples. Both are defined as the inverse probability of selection, but the methods of selection differ, therefore the probabilities differ. For the RDD frame, the selection of telephone numbers is done within each RDD stratum. An RDD stratum is an aggregation of area code prefixes (ACP: the first six digits of a 10-digit telephone number), with each ACP containing valid banks of one hundred numbers (see Norris and Paton[8] for more details). Therefore, the probability of selection is the ratio between the number of sampled units and one hundred times the number of banks within the RDD stratum.

For the list frame, telephone numbers are selected among all numbers available on the list that are within the specific HR. Hence, the probability of selection corresponds to the ratio of the number of sampled units to the number of telephone numbers on the list within the HR. Since sampling for the telephone frame was done on a monthly basis (see adjustment T1) and because the list frame was updated during the survey, the number of available telephone numbers within each HR may have changed from one month to another modifying the probability of selection during the survey. The inverse of these probabilities represents the initial weight T0.

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

T1 – Number of months

Contrary to the area frame, where the entire sample was selected at the beginning of the sampling process, samples were drawn on a monthly basis for the telephone frame. Each of these monthly samples came with an initial weight that made each sample representative at the HR level. However, to ensure that the total sample would represent the population only once, an adjustment factor had to be applied to reduce the weights of each monthly sample. The adjustment factor applied to each monthly sample was equal to the proportion of the total sample represented by the each monthly sample. Since the coverage is different from one version of the list frame to another, the adjustment was done separately for the two versions which means that the sample of each version represented the total population. Thus, at this point, the total list frame sample represents about two times the total population. To correct this situation, samples from the two versions were later combined (in step T4) in such a way that the list frame would represent the total population only once. Therefore, the weight T1 was obtained by multiplying the initial weight T0 by the factor defined above.

T2 - Removal of out-of-scope numbers

Telephone numbers leading to businesses, institutions or other out-of-scope dwellings, as well as numbers not in service or any other non-working numbers are all examples of out-of-scope cases for the telephone frame. As was done for the area frame, these cases were simply removed from the process, leaving only in-scope dwellings in the sample. These in-scope dwellings kept the same weight as in the previous step, now called weight T2.

T3 – Coverage of the list frames

Since the list frame does not cover some phone numbers, which are actually covered by the RDD frame, an adjustment had to be applied to the initial weights of the list frame units to make both frames comparable in terms of coverage. The adjustment consisted of inflating the weights of the list frame units by the amount of undercoverage, individually for each HR. Estimating the undercoverage was a challenging task and was done using the data collected from the CCHS Cycle 3.1 area frame sample. For all people interviewed via the area frame, the questionnaire included a set of questions verifying if the household had a telephone, how many residential lines it had, and the phone number for each line. The desired coverage rate was derived by simply computing the percentage of all collected numbers that were present in the list frame. The coverage rate was computed independently within each version of the list frame since the coverage varies from one version to another. The inverse of this rate represents the factor used for this adjustment. The factor, once multplied by the weight T2, resulted in the weight T3.

T4 - Combination of the list frames

Up to this step, samples from the two versions of the list frame represent the entire population of the HR where the list frame was used. That is to say that the population is represented twice with the weights up to this point. They both had to be combined so that together they would represent the total population only once. An adjustment factor based on size of samples used in each version was computed as follows:

$$\frac{Size\ of\ the\ sample\ coming\ from\ the\ version\ of\ the\ list\ frame}{Size\ of\ the\ total\ sample\ coming\ from\ the\ list\ frame}$$

These factors were calculated and applied independently within each HR where the list frame was used. This adjustment was equal to 1 for the HRs where the RDD frame was used. Consequently, the weight T4 was obtained by multiplying the weight T3 by the combining factor.

T5 – Household nonresponse

The adjustment applied here to compensate for the effect of household nonresponse is identical to the one applied for the area frame (adjustment A4). As was done for A4, the collection period was a significant characteristic for explaining the nonresponse. That variable was hence used to define the adjustment classes. The adjustment factor calculated within each class was obtained as follows:

$$\frac{Sum\ of\ weights\ T4\ for\ all\ households}{Sum\ of\ weights\ T4\ for\ all\ responding\ households}$$

The weight T4 of responding households was multiplied by this factor to produce the weight T5. Non-responding households are removed from the process at this point.

T6 - Households without telephone

A certain proportion of the Canadian population does not have access to a private residential telephone line. As explained in step T3, information about the presence of a telephone was collected for the area frame sample, which was used here to estimate the proportion of households without a phone line at the HR level. Similar to T3, the telephone frame sample weights were inflated based on proportions observed using the area frame data, adjusting the weights for the under-coverage of the frame for this uncovered sub-population. The factor used for this adjustment corresponded to the inverse of the estimated proportion, and once multiplied by the weight T5, resulted in weight T6.

T7 – Creation of person level weight

As was done in adjustment A5, this adjustment converts the household level weight to a person level weight. Since the algorithm of selection of the person within the household is the same as the one used for the area frame, computation of the adjustment factor was done the same way. This factor, multiplied by the weight T6, gave the weight T7.

T8 - Person nonresponse

This adjustment was similar to the adjustment A6 used for the area frame. It consisted of compensating for the effect of nonresponse at the person level. As was done for A6, an approach based on adjustment classes was used, where classes were defined from variables available for all selected persons, respondent or not (see A6 for the list of variables available). Within each class, an adjustment factor was calculated as follows:

$$\frac{\textit{Sum of weights } T7 \textit{ for all selected persons}}{\textit{Sum of weights } T7 \textit{ for all responding selected persons}}$$

The weight T7 of responding persons was therefore multiplied by this adjustment factor to produce the weight T8. Non-responding persons were dropped out of the weighting process at this point.

T9 – Multiple lines

Some households can possess more than one residential telephone line. This has an impact on the weighting: having more lines translates into having a higher probability of being selected. Therefore, the weights needed to be adjusted for the number of residential telephone lines the household had. Even though this characteristic is relative to the household, the information is only collected during the interview of the selected person. This is why the adjustment is applied at this stage of the weighting. The adjustment factor represented the inverse of the number of lines in the household. The weight T9 was therefore obtained by multiplying this factor by the weight T8.

Since this adjustment was the last one for the sample drawn from the telephone frame, the weight T9 represents the ***final telephone frame weight***. This weight was later integrated, in step I1, with the final area frame weight to finally create the final CCHS Cycle 3.1 weight.

8.1.3 Integration of the area and telephone frames (I1)

This step consisted in integrating the final area and telephone frame sampling weights into a single weight by applying a method of integration[9]. An adjustment factor between 0 and 1 was determined at the health region level in such a way that it represented the relative importance of

[9] Skinner, C.J. and Rao, J.N.K. 1996. Estimation in Dual Frame Surveys with Complex Designs. *Journal of the American Statistical Association*. 91, 433, 349-356.

each sample in the total sample. This relative importance was measured in terms of sample size and design effect. For the sample size, the higher the proportion that the sample represented of the total sample, the higher its relative importance. For the design effect, the relative importance was larger for units coming from the frame that had the smallest design effect. To obtain the integration adjustment factor, a factor α was first calculated as follows:

$$\alpha = \frac{n_A}{R} \Big/ \left(\frac{n_A}{R} + n_T \right)$$

where n_A and n_T represent the area and telephone frames sample sizes respectively, while R represents the median ratio of the design effects for several key characteristics estimated for each frame. The weight of the area frame units was multiplied by this factor α, while the weight of the telephone frame units was multiplied by 1- α. Note that in the case where a HR was covered by only one frame, the adjustment factor was equal to 1. The product between the factor derived here and the final weight calculated earlier (A6 or T9 depending on which frame the unit belongs to), gave the integrated weight I1.

8.1.4 Seasonal effect and Winsorization (I2)

For CCHS Cycle 3.1, the initial plan was to allocate the data collection equally throughout twelve months of the survey's reference year, partly to control for the seasonal effect in the data collected. However, some events affected these plans, with the result that an additional adjustment had to be added to ensure that there was no seasonal effect in the estimates produced. The adjustment applied in I2 was done so that the sum of the weights of all units interviewed during one of the four seasons would represent exactly 25 % of the total sum of weights. In other words, after applying the adjustment, the portion of the sample interviewed each season represented 25 % of the total population for each HR.

The four seasons defined for the CCHS Cycle 3.1 are the periods covering September to November, December to February, March to May, and June to August. The adjustment factor I2 used to control the seasonal effect for a person interviewed during season *S*, is defined as:

$$\frac{\textit{Sum of weights I1 for the total sample}}{4 \times \textit{sum of weights I1 for the sample interviewed during season S}}$$

This seasonal adjustment applied to the weight I1 results in the weight I2.

Winsorization

Note that following the series of adjustments applied to the respondents, some units may come out with outlier weights compared with other units of the same HR. Some respondents could represent a large proportion of their HR and hence strongly influence estimates for their HR. In order to prevent this, the weight of the outlier units that represent a large proportion of their HR-age-sex group is adjusted downward using a "winsorization" trimming approach.

8.1.5 Post-stratification (I3)

The final step necessary to obtain the final CCHS Cycle 3.1 weight is post-stratification. Post-stratification is done to ensure that the sum of the final weights corresponds to the population estimates defined at the HR level, for all 10 age-sex groups of interest, that is, the five age groups 12-19, 20-29, 30-44, 45-64, 65+, for both males and females. Note that the post-stratification was done using a revised geography that contained 4 regions instead of the original 6 used at the design stage and throughout data collection for Newfoundland and Labrador and for Ontario there was one less health region. Note also that the post-stratification was done at the CLSC region level for the three Quebec regions (2401, 2406, and 2413) that paid to get extra sample.

The 2005 population estimates were based on 2001 Census counts and counts of birth, death, immigration and emigration since that time. The average of these 2005 monthly estimates for each of the HR-age-sex post-strata was used to post-stratify. The weight I2 was therefore adjusted to obtain the final weight I3 with the help of the adjustment factor I3 defined as follows:

$$\frac{\textit{Population estimate for the HR - age - sex group of the respondent}}{\textit{Sum of weights I2 for the HR - age - sex group of the respondent}}$$

Consequently, the weight I3 corresponds to the ***final CCHS Cycle 3.1 weight*** that can be found on the data file with the variable name WTSE_M.

8.1.6 Particular aspects of the weighting in the three territories

As described in Section 5, the sampling frame used in the three territories was somewhat different from the one used in the provinces. Therefore, the weighting strategy had to be adapted to comply with these differences. This section summarises the changes applied to the steps described in sub-sections 8.1.1 to 8.1.5.

For the area frame, as mentioned in sub-section 5.4.1, an additional stage of selection was added in the territories where each territory was initially stratified into groupings of communities and one community was selected within each group. Note that the capital of each territory formed a stratum on its own and was consequently selected automatically at this first stage. This had an effect in the computation of the probability of selection, and therefore in the value of the initial weight (A0). Once the initial weight was calculated, the same series of adjustments (A1 to A6) was applied to the area frame units. Household-level and person-level nonresponse adjustment classes were built in the same way as for the provinces, using the same set of variables available.

For the weighting of the telephone frame units, it should first be noted that only the RDD frame was used for the territories, and exclusively in the Yukon and Northwest Territories capitals. Consequently, this eliminated the need of adjustments T3 (coverage of the list frames) and T4 (combination of the list frames). All other adjustments were applied. Finally, adjustment T6 (household without telephone lines) was also subject to a slight modification since the RDD frame was used only in the capitals. The proportions of households without telephone lines were derived,

as was done for the provinces, using the area frame data, but by excluding the data from households located outside the capitals from the calculations.

The two sets of weights (area and telephone) were subsequently integrated, then adjusted for the seasonal effect, and finally post-stratified in a similar way to what was done for the provinces, with the exception of two details. First, the integration was applied only to units located in the Yukon and Northwest Territories capitals since the other communities were covered only by the area frame. As well, for Nunavut, the population counts used for calibration only represent 70% of the entire population because of the under-coverage of the area frame that was described in section 5.4.1.

9. Data quality

9.1 Response rates

In total and after removing the out-of-scope units, 168,464 households were selected to participate in the CCHS Cycle 3.1. Out of these selected households a response was obtained for 143,076 of them, which results in an overall household-level response rate of 84.9%. Among these responding households, 143,076 individuals (one per household) were selected to participate in the CCHS Cycle 3.1, out of which a response was obtained for 132,947 individuals, resulting in an overall person-level response rate of 92.9%. At the Canada level, this yields a combined response rate of 78.9% for the CCHS Cycle 3.1. Table 9.1 provides combined response rates as well as relevant information for their calculation by health region or combined health region.

Next, we describe how the various components of the equation should be handled to correctly compute combined response rates.

Household-level response rate

$$\text{HHRR} = \frac{\textit{\# of responding households in both frames}}{\textit{all in-scope households in both frames}}$$

Person-level response rate

$$\text{PPRR} = \frac{\textit{\# of responding persons in both frames}}{\textit{all selected persons in both frames}}$$

Combined response rate = HHRR x PPRR

Next is an example on how to calculate the combined response rate for Canada using the information found in Table 9.1.

$$\text{HHRR} = \frac{69\,417 + 73\,659}{78\,396 + 90\,068} = \frac{143\,076}{168\,464} = 0.849$$

$$\text{PPRR} = \frac{65\,039 + 67\,908}{69\,417 + 73\,659} = \frac{132\,947}{143\,076} = 0.929$$

Combined response rate = 0.849 x 0.929

= 0.789

= **78.9%**

Table 9.1 Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases / Tous les cas
Prov./ Terr Prov./ Terr.	Health Region Région socio-sanitaire	# in scope HH # mén. cibles	# resp. HH # mén. rép.	HH resp. rates Taux de rép. mén.	# pers. select. # pers. sélect.	# resp. # rép.	Pers. resp. rates Taux de rép. pers.	Resp. rates Taux de rép.	# in scope HH # mén. cibles	# resp. HH # mén. rép.	HH resp. rates Taux de rép. mén.	# pers. select. # pers. sélect.	# resp. # rép.	Pers. resp. rates Taux de rép. pers.	Resp. rates Taux de rép.	Combined resp. rates Taux de rép. combiné
CA	Total	78396	69417	88.5	69417	65039	93.7	83.0	89040	72850	81.8	72850	67182	92.2	75.5	79.0
NL	Total	2283	2112	92.5	2112	2001	94.7	87.6	2516	2271	90.3	2271	2110	92.9	83.9	85.7
	10911	1050	966	92.0	966	917	94.9	87.3	1076	979	91.0	979	897	91.6	83.4	85.3
	10912	416	383	92.1	383	362	94.5	87.0	465	420	90.3	420	385	91.7	82.8	84.8
	10913*	817	763	93.4	763	722	94.6	88.4	975	872	89.4	872	828	95.0	84.9	86.5
PE	Total	1157	1049	90.7	1049	995	94.9	86.0	1279	1120	87.6	1120	1036	92.5	81.0	83.4
	11901*	691	636	92.0	636	608	95.6	88.0	851	749	88.0	749	687	91.7	80.7	84.0
	11903	466	413	88.6	413	387	93.7	83.0	428	371	86.7	371	349	94.1	81.5	82.3
NS	Total	2944	2607	88.6	2607	2417	92.7	82.1	3098	2806	90.6	2806	2649	94.4	85.5	83.8
	12901	438	409	93.4	409	378	92.4	86.3	495	450	90.9	450	422	93.8	85.3	85.7
	12902	408	369	90.4	369	350	94.9	85.8	345	317	91.9	317	306	96.5	88.7	87.1
	12903	414	378	91.3	378	350	92.6	84.5	439	396	90.2	396	368	92.9	83.8	84.2
	12904	363	330	90.9	330	312	94.5	86.0	479	431	90.0	431	401	93.0	83.7	84.7
	12905	481	429	89.2	429	394	91.8	81.9	512	458	89.5	458	442	96.5	86.3	84.2
	12906	840	692	82.4	692	633	91.5	75.4	828	754	91.1	754	710	94.2	85.7	80.5
NB	Total	2982	2706	90.7	2706	2542	93.9	85.2	3117	2748	88.2	2748	2558	93.1	82.1	83.6
	13901	599	525	87.6	525	493	93.9	82.3	572	506	88.5	506	475	93.9	83.0	82.7
	13902	558	501	89.8	501	469	93.6	84.1	616	550	89.3	550	511	92.9	83.0	83.5
	13903	567	511	90.1	511	486	95.1	85.7	558	486	87.1	486	461	94.9	82.6	84.2
	13904*	558	518	92.8	518	487	94.0	87.3	642	563	87.7	563	519	92.2	80.8	83.8
	13906*	700	651	93.0	651	607	93.2	86.7	729	643	88.2	643	592	92.1	81.2	83.9
QC	Total	14674	12895	87.9	12895	12124	94.0	82.6	23498	18973	80.7	18973	17041	89.8	72.5	76.4
	24901	840	785	93.5	785	765	97.5	91.1	3579	3012	84.2	3012	2753	91.4	76.9	79.6
	24902	745	674	90.5	674	645	95.7	86.6	737	638	86.6	638	585	91.7	79.4	83.0
	24903	1181	1020	86.4	1020	973	95.4	82.4	1219	1003	82.3	1003	910	90.7	74.7	78.5
	24904	983	883	89.8	883	851	96.4	86.6	1101	904	82.1	904	831	91.9	75.5	80.7
	24905	827	701	84.8	701	655	93.4	79.2	765	646	84.4	646	578	89.5	75.6	77.4
	24906	2017	1698	84.2	1698	1570	92.5	77.8	5793	4346	75.0	4346	3820	87.9	65.9	69.0
	24907	793	663	83.6	663	626	94.4	78.9	831	692	83.3	692	630	91.0	75.8	77.3
	24908	619	570	92.1	570	530	93.0	85.6	799	687	86.0	687	640	93.2	80.1	82.5
	24909	763	665	87.2	665	633	95.2	83.0	702	603	85.9	603	542	89.9	77.2	80.2
	24911	701	656	93.6	656	619	94.4	88.3	784	654	83.4	654	595	91.0	75.9	81.8
	24912	854	776	90.9	776	736	94.8	86.2	927	763	82.3	763	678	88.9	73.1	79.4
	24913	950	832	87.6	832	723	86.9	76.1	2454	1941	79.1	1941	1709	88.0	69.6	71.4
	24914	914	815	89.2	815	773	94.8	84.6	934	762	81.6	762	674	88.5	72.2	78.3
	24915	939	803	85.5	803	757	94.3	80.6	1069	841	78.7	841	775	92.2	72.5	76.3

Table 9.1 Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases / Tous les cas
Prov./ Terr Prov./ Terr.	Health Region Région socio-sanitaire	# in scope HH # mén. cibles	# resp. HH # mén. rép.	HH resp. rates Taux de rép. mén.	# pers. select. # pers. sélect.	# resp. # rép.	Pers. resp. rates Taux de rép. pers.	Resp. rates Taux de rép.	# in scope HH # mén. cibles	# resp. HH # mén. rép.	HH resp. rates Taux de rép. mén.	# pers. select. # pers. sélect.	# resp. # rép.	Pers. resp. rates Taux de rép. pers.	Resp. rates Taux de rép.	Combined resp. rates Taux de rép. combiné
	24916	1548	1354	87.5	1354	1268	93.6	81.9	1804	1481	82.1	1481	1321	89.2	73.2	77.2
ON	Total	25781	22457	87.1	22457	20869	92.9	80.9	28458	22823	80.2	22823	20897	91.6	73.4	77.0
	35926	453	359	79.2	359	334	93.0	73.7	662	526	79.5	526	497	94.5	75.1	74.5
	35927	492	425	86.4	425	390	91.8	79.3	551	447	81.1	447	408	91.3	74.0	76.5
	35930	978	831	85.0	831	759	91.3	77.6	1110	879	79.2	879	820	93.3	73.9	75.6
	35931	467	406	86.9	406	382	94.1	81.8	448	353	78.8	353	326	92.4	72.8	77.4
	35933	634	556	87.7	556	530	95.3	83.6	600	509	84.8	509	481	94.5	80.2	81.9
	35934	451	382	84.7	382	355	92.9	78.7	520	413	79.4	413	374	90.6	71.9	75.1
	35935	600	515	85.8	515	479	93.0	79.8	654	552	84.4	552	511	92.6	78.1	78.9
	35936	820	732	89.3	732	689	94.1	84.0	969	793	81.8	793	710	89.5	73.3	78.2
	35937	1145	961	83.9	961	875	91.1	76.4	1156	885	76.6	885	807	91.2	69.8	73.1
	35938	522	477	91.4	477	437	91.6	83.7	629	505	80.3	505	471	93.3	74.9	78.9
	35939*	738	655	88.8	655	623	95.1	84.4	760	637	83.8	637	583	91.5	76.7	80.5
	35940	435	418	96.1	418	404	96.7	92.9	515	423	82.1	423	390	92.2	75.7	83.6
	35941	645	543	84.2	543	506	93.2	78.4	636	528	83.0	528	495	93.8	77.8	78.1
	35942	469	430	91.7	430	392	91.2	83.6	630	494	78.4	494	448	90.7	71.1	76.4
	35943	634	562	88.6	562	518	92.2	81.7	615	525	85.4	525	479	91.2	77.9	79.8
	35944	922	750	81.3	750	692	92.3	75.1	978	792	81.0	792	726	91.7	74.2	74.6
	35945	334	289	86.5	289	274	94.8	82.0	467	393	84.2	393	367	93.4	78.6	80.0
	35946	957	861	90.0	861	822	95.5	85.9	1005	805	80.1	805	722	89.7	71.8	78.7
	35947*	701	649	92.6	649	607	93.5	86.6	745	644	86.4	644	595	92.4	79.9	83.1
	35949	381	324	85.0	324	301	92.9	79.0	420	346	82.4	346	323	93.4	76.9	77.9
	35951	1228	1017	82.8	1017	930	91.4	75.7	1357	1122	82.7	1122	1045	93.1	77.0	76.4
	35952	462	409	88.5	409	368	90.0	79.7	467	402	86.1	402	385	95.8	82.4	81.1
	35953	1356	1147	84.6	1147	1055	92.0	77.8	1596	1203	75.4	1203	1070	88.9	67.0	72.0
	35955	553	471	85.2	471	430	91.3	77.8	540	456	84.4	456	416	91.2	77.0	77.4
	35956	399	362	90.7	362	334	92.3	83.7	477	389	81.6	389	359	92.3	75.3	79.1
	35957	475	434	91.4	434	408	94.0	85.9	451	371	82.3	371	345	93.0	76.5	81.3
	35958	621	557	89.7	557	524	94.1	84.4	659	549	83.3	549	508	92.5	77.1	80.6
	35960	824	730	88.6	730	689	94.4	83.6	876	710	81.1	710	644	90.7	73.5	78.4
	35961	646	589	91.2	589	558	94.7	86.4	713	607	85.1	607	560	92.3	78.5	82.3
	35962	567	516	91.0	516	482	93.4	85.0	631	540	85.6	540	494	91.5	78.3	81.5
	35965	999	865	86.6	865	780	90.2	78.1	1015	841	82.9	841	781	92.9	76.9	77.5
	35966	646	590	91.3	590	555	94.1	85.9	730	599	82.1	599	548	91.5	75.1	80.2
	35968	870	779	89.5	779	742	95.3	85.3	1014	808	79.7	808	732	90.6	72.2	78.2
	35970	1114	989	88.8	989	924	93.4	82.9	1354	1016	75.0	1016	900	88.6	66.5	73.9

Table 9.1 Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases / Tous les cas
Prov./ Terr Prov./ Terr.	Health Region Région socio-sanitaire	# in scope HH # mén. cibles	# resp. HH # mén. rép.	HH resp. rates Taux de rép. mén.	# pers. select. # pers. sélect.	# resp. # rép.	Pers. resp. rates Taux de rép. pers.	Resp. rates Taux de rép.	# in scope HH # mén. cibles	# resp. HH # mén. rép.	HH resp. rates Taux de rép. mén.	# pers. select. # pers. sélect.	# resp. # rép.	Pers. resp. rates Taux de rép. pers.	Resp. rates Taux de rép.	Combined resp. rates Taux de rép. combiné
	35995	2243	1877	83.7	1877	1721	91.7	76.7	2508	1761	70.2	1761	1577	89.6	62.9	69.4
MB	Total	4378	3973	90.7	3973	3785	95.3	86.5	4443	3758	84.6	3758	3567	94.9	80.3	83.3
	46910	1231	1055	85.7	1055	993	94.1	80.7	1297	1087	83.8	1087	1037	95.4	80.0	80.3
	46915*	694	640	92.2	640	617	96.4	88.9	755	639	84.6	639	607	95.0	80.4	84.5
	46920*	762	710	93.2	710	667	93.9	87.5	606	521	86.0	521	479	91.9	79.0	83.8
	46930	422	387	91.7	387	365	94.3	86.5	371	316	85.2	316	306	96.8	82.5	84.6
	46940	456	425	93.2	425	407	95.8	89.3	453	392	86.5	392	376	95.9	83.0	86.1
	46960*	813	756	93.0	756	736	97.4	90.5	961	803	83.6	803	762	94.9	79.3	84.4
SK	Total	4431	4053	91.5	4053	3883	95.8	87.6	4804	4080	84.9	4080	3882	95.1	80.8	84.1
	47901*	1074	999	93.0	999	965	96.6	89.9	989	828	83.7	828	783	94.6	79.2	84.7
	47904	812	707	87.1	707	679	96.0	83.6	674	576	85.5	576	549	95.3	81.5	82.6
	47905*	590	548	92.9	548	524	95.6	88.8	764	643	84.2	643	620	96.4	81.2	84.5
	47906	860	778	90.5	778	738	94.9	85.8	705	591	83.8	591	567	95.9	80.4	83.4
	47907*	689	643	93.3	643	608	94.6	88.2	651	573	88.0	573	543	94.8	83.4	85.9
	47909*	406	378	93.1	378	369	97.6	90.9	1021	869	85.1	869	820	94.4	80.3	83.3
AB	Total	7139	6388	89.5	6388	5970	93.5	83.6	7331	6145	83.8	6145	5830	94.9	79.5	81.5
	48920	595	551	92.6	551	530	96.2	89.1	615	524	85.2	524	501	95.6	81.5	85.2
	48921	507	470	92.7	470	450	95.7	88.8	446	375	84.1	375	363	96.8	81.4	85.3
	48922	1658	1409	85.0	1409	1308	92.8	78.9	1691	1405	83.1	1405	1340	95.4	79.2	79.1
	48923	786	713	90.7	713	654	91.7	83.2	812	682	84.0	682	649	95.2	79.9	81.5
	48924	524	485	92.6	485	461	95.1	88.0	519	455	87.7	455	423	93.0	81.5	84.8
	48925	1576	1395	88.5	1395	1291	92.5	81.9	1565	1312	83.8	1312	1252	95.4	80.0	81.0
	48926	559	518	92.7	518	491	94.8	87.8	670	556	83.0	556	528	95.0	78.8	82.9
	48927*	934	847	90.7	847	785	92.7	84.0	1013	836	82.5	836	774	92.6	76.4	80.1
BC	Total	9687	8513	87.9	8513	7976	93.7	82.3	10247	7925	77.3	7925	7431	93.8	72.5	77.3
	59911	348	332	95.4	332	321	96.7	92.2	324	278	85.8	278	258	92.8	79.6	86.2
	59912	328	305	93.0	305	292	95.7	89.0	377	303	80.4	303	287	94.7	76.1	82.1
	59913	679	629	92.6	629	607	96.5	89.4	714	559	78.3	559	519	92.8	72.7	80.8
	59914	553	483	87.3	483	443	91.7	80.1	601	482	80.2	482	457	94.8	76.0	78.0
	59921	591	533	90.2	533	506	94.9	85.6	660	513	77.7	513	486	94.7	73.6	79.3
	59922	916	851	92.9	851	816	95.9	89.1	968	727	75.1	727	669	92.0	69.1	78.8
	59923	956	844	88.3	844	792	93.8	82.8	1069	802	75.0	802	741	92.4	69.3	75.7
	59931	500	425	85.0	425	403	94.8	80.6	575	400	69.6	400	379	94.8	65.9	72.7
	59932	1085	822	75.8	822	780	94.9	71.9	1134	797	70.3	797	735	92.2	64.8	68.3
	59933	583	518	88.9	518	478	92.3	82.0	723	571	79.0	571	543	95.1	75.1	78.2

Table 9.1 Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases / Tous les cas
Prov./ Terr Prov./ Terr.	Health Region Région socio-sanitaire	# in scope HH # mén. cibles	# resp. HH # mén. rép.	HH resp. rates Taux de rép. mén.	# pers. select. # pers. sélect.	# resp. # rép.	Pers. resp. rates Taux de rép. pers.	Resp. rates Taux de rép.	# in scope HH # mén. cibles	# resp. HH # mén. rép.	HH resp. rates Taux de rép. mén.	# pers. select. # pers. sélect.	# resp. # rép.	Pers. resp. rates Taux de rép. pers.	Resp. rates Taux de rép.	Combined resp. rates Taux de rép. combiné
	59941	853	759	89.0	759	691	91.0	81.0	881	694	78.8	694	660	95.1	74.9	77.9
	59942	678	605	89.2	605	572	94.5	84.4	662	538	81.3	538	513	95.4	77.5	81.0
	59943	277	245	88.4	245	235	95.9	84.8	355	290	81.7	290	279	96.2	78.6	81.3
	59951*	834	702	84.2	702	650	92.6	77.9	693	545	78.6	545	505	92.7	72.9	75.6
	590052	506	460	90.9	460	390	84.8	77.1	511	426	83.4	426	400	93.9	78.3	77.7
Terr.	Total	2940	2664	90.6	2664	2477	93.0	84.3	249	201	80.7	201	181	90.0	72.7	83.3

* = collapsed health regions

9.2 Survey Errors

The estimates derived from this survey are based on a sample of individuals. Somewhat different figures might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. than those actually used. The difference between the estimates obtained from the sample and the results from a complete count under similar conditions is called the *sampling error* of the estimate.

Errors which are not related to sampling may occur at almost every phase of a survey operation. Interviewers may misunderstand instructions, respondents may make errors in answering questions, the answers may be incorrectly entered on the computer and errors may be introduced in the processing and tabulation of the data. These are all examples of *non-sampling errors*.

9.2.1 Non-sampling Errors

Over a large number of observations, randomly occurring errors will have little effect on estimates derived from the survey. However, errors occurring systematically will contribute to biases in the survey estimates. Considerable time and effort was made to reduce non-sampling errors in the CCHS Cycle 3.1. Quality assurance measures were implemented at each step of data collection and processing to monitor the quality of the data. These measures included the use of highly skilled interviewers, extensive training with respect to the survey procedures and questionnaire, and the observation of interviewers to detect problems. Testing of the CAI application and field tests were also essential procedures to ensure that data collection errors were minimized.

A major source of non-sampling errors in surveys is the effect of *non-response* on the survey results. The extent of non-response varies from partial non-response (failure to answer just one or some questions) to total non-response. Partial non-response to the CCHS Cycle 3.1 was minimal; once the questionnaire was started, it tended to be completed with very little non-response. Total non-response occurred either because a person refused to participate in the survey or because the interviewer was unable to contact the selected person. Total non-response was handled by adjusting the weight of persons who responded to the survey to compensate for those who did not respond. See Section 8 for details on the weight adjustment for non-response.

9.2.2 Sampling Errors

Since it is an unavoidable fact that estimates from a sample survey are subject to sampling error, sound statistical practice calls for researchers to provide users with some indication of the magnitude of this sampling error. The basis for measuring the potential size of sampling errors is the standard deviation of the estimates derived from survey results. However, because of the large variety of estimates that can be produced from a survey, the standard deviation of an estimate is usually expressed relative to the estimate to which it pertains. This resulting measure, known as the coefficient of variation (CV) of an estimate, is obtained by dividing the standard deviation of the estimate by the estimate itself and is expressed as a percentage of the estimate.

For example, suppose hypothetically that it is estimated that 25% of Canadians aged 12 and over are regular smokers and that this estimate is found to have a standard deviation of 0.003. Then the CV of the estimate is calculated as:

$$(0.003/0.25) \times 100\% = 1.20\%$$

Statistics Canada commonly uses CV results when analyzing data and urges users producing estimates from the CCHS Cycle 3.1 data files to also do so. For details on how to determine CVs, see Section 11. For guidelines on how to interpret CV results, see the table at the end of Sub-section 10.4.

10. Guidelines for tabulation, analysis and release

This section of the documentation outlines the guidelines to be adhered to by users tabulating, analyzing, publishing or otherwise releasing any data derived from the survey PUMF. With the aid of these guidelines, users of microdata should be able to produce figures that are in close agreement with those produced by Statistics Canada and, at the same time, will be able to develop currently unpublished figures in a manner consistent with these established guidelines.

10.1 Rounding guidelines

In order that estimates for publication or other release derived from this PUMF correspond to those produced by Statistics Canada, users are urged to adhere to the following guidelines regarding the rounding of such estimates:

a) Estimates in the main body of a statistical table are to be rounded to the nearest hundred units using the normal rounding technique. In normal rounding, if the first or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is raised by one. For example, in normal rounding to the nearest 100, if the last two digits are between 00 and 49, they are changed to 00 and the preceding digit (the hundreds digit) is left unchanged. If the last digits are between 50 and 99 they are changed to 00 and the proceeding digit is incremented by 1;

b) Marginal sub-totals and totals in statistical tables are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest 100 units using normal rounding;

c) Averages, proportions, rates and percentages are to be computed from unrounded components (i.e., numerators and/or denominators) and then are to be rounded themselves to one decimal using normal rounding. In normal rounding to a single digit, if the final or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is increased by 1;

d) Sums and differences of aggregates (or ratios) are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest 100 units (or the nearest one decimal) using normal rounding;

e) In instances where, due to technical or other limitations, a rounding technique other than normal rounding is used resulting in estimates to be published or otherwise released that differ from corresponding estimates published by Statistics Canada, users are urged to note the reason for such differences in the publication or release document(s);

f) Under no circumstances are unrounded estimates to be published or otherwise released by users. Unrounded estimates imply greater precision than actually exists.

10.2 Sample weighting guidelines for tabulation

The sample design used for this survey was not self-weighting. That is to say, the sampling weights are not identical for all individuals in the sample. When producing simple estimates, including the production of ordinary statistical tables, users must apply the proper sampling weight. If proper weights are not used, the estimates derived from the PUMF cannot be considered to be representative of the survey population, and will not correspond to those produced by Statistics Canada.

Users should also note that some software packages might not allow the generation of estimates that exactly match those available from Statistics Canada, because of their treatment of the weight field.

10.2.1 Definitions: categorical estimates, quantitative estimates

Before discussing how the survey data can be tabulated and analyzed, it is useful to describe the two main types of point estimates of population characteristics that can be generated from the PUMF.

Categorical estimates:

Categorical estimates are estimates of the number or percentage of the surveyed population possessing certain characteristics or falling into some defined category. The number of individuals who smoke daily is an example of such an estimate. An estimate of the number of persons possessing a certain characteristic may also be referred to as an estimate of an aggregate.

Example of categorical question:

> *At the present do/does ... smoke cigarettes daily, occasionally or not at all?* ***(SMKE_202)***
>
> __ Daily
> __ Occasionally
> __ Not at all

Quantitative estimates:

Quantitative estimates are estimates of totals or of means, medians and other measures of central tendency of quantities based upon some or all of the members of the surveyed population.

An example of a quantitative estimate is the average number of cigarettes smoked per day by individuals who smoke daily. The numerator is an estimate of the total number of cigarettes smoked per day by individuals who smoke daily, and its denominator is an estimate of the number of individuals who smoke daily.

Example of quantitative question:

How many cigarettes do/does you/he/she smoke each day now? ***(SMKE_204)***
|_|_| Number of cigarettes

10.2.2 Tabulation of categorical estimates

Estimates of the number of people with a certain characteristic can be obtained from the PUMF by summing the final weights of all records possessing the characteristic of interest.

Proportions and ratios of the form $\hat{X} / \hat{Y}$ are obtained by:

a) summing the final weights of records having the characteristic of interest for the numerator ($\hat{X}$);

b) summing the final weights of records having the characteristic of interest for the denominator ($\hat{Y}$); then

c) dividing the numerator estimate by the denominator estimate.

10.2.3 Tabulation of quantitative estimates

Estimates of sums or averages for quantitative variables can be obtained using the following three steps (only step a) is necessary to obtain the estimate of a sum):

a) multiplying the value of the variable of interest by the final weight and summing this quantity over all records of interest to obtain the numerator ($\hat{X}$);

b) summing the final weights of records having the characteristic of interest for the denominator ($\hat{Y}$); then

c) dividing the numerator estimate by the denominator estimate.

For example, to obtain the estimate of the average number of cigarettes smoked each day by individuals who smoke daily, first compute the numerator ($\hat{X}$) by summing the product between the value of variable **SMKE_204** and the weight **WTSE_M.** Next, sum this value over those records with a value of "daily" to the variable **SMKE_202**. The denominator ($\hat{Y}$) is obtained by summing the final weight of those records with a value of "daily" to the variable **SMKE_202**. Divide ($\hat{X}$) by ($\hat{Y}$) to obtain the average number of cigarettes smoked each day by daily smokers.

10.3 Guidelines for statistical analysis

The CCHS is based upon a complex design, with stratification and multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures can differ from what is appropriate in a sample survey framework, with the result that while in many cases the estimates produced by the packages are correct, the variances that are calculated are almost meaningless.

For many analysis techniques (for example linear regression, logistic regression, analysis of variance), a method exists that can make the application of standard packages more meaningful. If the weights on the records are rescaled so that the average weight is one (1), then the results produced by the standard packages will be more reasonable; they still will not take into account the stratification and clustering of the sample's design, but they will take into account the unequal probabilities of selection. The rescaling can be accomplished by using in the analysis a weight equal to the original weight divided by the average of the original weights for the sampled units (people) contributing to the estimator in question.

In order to provide a means of assessing the quality of tabulated estimates, Statistics Canada has produced a set of Approximate Coefficients of Variations Tables (commonly referred to as "CV Tables") for the CCHS. These tables can be used to obtain approximate coefficients of variation for categorical-type estimates and proportions. See Section 11 for more details.

10.4 Release guidelines

Before releasing and/or publishing any estimate from the PUMF, users must first determine the number of sampled respondents having the characteristic of interest (for example, the number of respondents who smoke when interested in the proportion of smokers for a given population). If this number is less than 30, the unweighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. For weighted estimates based on sample sizes of 30 or more, users should determine the coefficient of variation of the rounded estimate and follow the guidelines below.

Table 10.1 Sampling variability guidelines

Type of Estimate	CV (in %)	Guidelines
Acceptable	$0.0 \leq CV \leq 16.5$	Estimates can be considered for general unrestricted release. Requires no special notation.
Marginal	$16.6 < CV \leq 33.3$	Estimates can be considered for general unrestricted release but should be accompanied by a warning cautioning subsequent users of the high sampling variability associated with the estimates. Such estimates should be identified by the letter E (or in some other similar fashion).
Unacceptable	$CV > 33.3$	Statistics Canada recommends not to release estimates of unacceptable quality. However, if the user chooses to do so then estimates should be flagged with the letter F (or in some other fashion) and the following warning should accompany the estimates: "The user is advised that . . .(specify the data) . . . do not meet Statistics Canada's quality standards for this statistical program. Conclusions based on these data will be unreliable and most likely invalid. These data and any consequent findings should not be published. If the user chooses to publish these data or findings, then this disclaimer must be published with the data."

11. Approximate sampling variability tables

In order to supply coefficients of variation that would be applicable to a wide variety of categorical estimates produced from this PUMF and that could be readily accessed by the user, a set of Approximate Sampling Variability Tables has been produced. These "look-up" tables allow the user to obtain an approximate coefficient of variation based on the size of the estimate calculated from the survey data.

The coefficients of variation (CV) are derived using the variance formula for simple random sampling and incorporating a factor which reflects the multi-stage, clustered nature of the sample design. This factor, known as the ***design effect,*** was determined by first calculating design effects for a wide range of characteristics and then choosing, for each table produced, a conservative value among all design effects relative to that table. The value chosen was then used to generate a table that applies to the entire set of characteristics.

The design effects, sample sizes and population counts used to produce the Approximate Sampling Variability Tables as well as the tables are presented in Appendix E. All coefficients of variation in the Approximate Sampling Variability Tables are approximate and, therefore, unofficial. Options concerning the computation of exact coefficients of variation are discussed in sub-section 11.7.

Remember: As indicated in Sampling Variability Guidelines in Section 10.4, if the number of observations on which an estimate is based is less than 10, the weighted estimate should not be released regardless of the value of the coefficient of variation. Coefficients of variation based on small sample sizes are too unpredictable to be adequately represented in the tables.

11.1 How to use the CV tables for categorical estimates

The following rules should enable the user to determine the approximate coefficients of variation from the Sampling Variability Tables for estimates of the number, proportion or percentage of the surveyed population possessing a certain characteristic and for ratios and differences between such estimates.

Rule 1: Estimates of numbers possessing a characteristic (aggregates)

The coefficient of variation depends only on the size of the estimate itself. On the appropriate Approximate Coefficients of Variations Table, locate the estimated number in the left-most column of the table (headed "Numerator of Percentage") and follow the asterisks (if any) across to the first figure encountered. Since not all the possible values for the estimate are available, the smallest value which is the closest must be taken (as an example, if the estimate is equal to 1,700 and the two closest available values are 1,000 and 2,000, the first has to be chosen). This figure is the approximate coefficient of variation.

Rule 2: Estimates of proportions or percentages of people possessing a characteristic

The coefficient of variation of an estimated proportion (or percentage) depends on both the size of the proportion and the size of the numerator upon which the proportion is based. Estimated proportions are relatively more reliable than the corresponding estimates of the numerator of the proportion when the proportion is based upon a sub-group of the population. This is due to the fact that the coefficients of variation of the latter type of estimates are based on the largest entry in a row of a particular table, whereas the coefficients of variation of the former type of estimators are based on some entry (not necessarily the largest) in that same row. (Note that in the tables the CVs decline in value reading across a row from left to right). For example, the estimated proportion of individuals who smoke daily out of those who smoke at all is more reliable than the estimated number who smoke daily.

When the proportion (or percentage) is based upon the total population covered by each specific table, the CV of the proportion is the same as the CV of the numerator of the proportion. In this case, this is equivalent to applying Rule 1.

When the proportion (or percentage) is based upon a subset of the total population (e.g., those who smoke at all), reference should be made to the proportion (across the top of the table) and to the numerator of the proportion (down the left side of the table). Since not all the possible values for the proportion are available, the smallest value which is the closest must be taken (for example, if the proportion is 23% and the two closest values available in the column are 20% and 25%, 20% must be chosen). The intersection of the appropriate row and column gives the coefficient of variation.

Rule 3: Estimates of differences between aggregates or percentages

The standard error of a difference between two estimates is approximately equal to the square root of the sum of squares of each standard error considered separately. That is, the standard error of a difference ($\hat{d} = \hat{X}_2 - \hat{X}_1$) is:

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

where $\hat{X}_1$ is estimate 1, $\hat{X}_2$ is estimate 2, and α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. The coefficient of variation of $\hat{d}$ is given by $\sigma_{\hat{d}} / \hat{d}$. This formula is accurate for the difference between independent populations or subgroups, but is only approximate otherwise. It will tend to overstate the error, if $\hat{X}_1$ and $\hat{X}_2$ are positively correlated and understate the error if $\hat{X}_1$ and $\hat{X}_2$ are negatively correlated.

Rule 4: Estimates of ratios

In the case where the numerator is a subset of the denominator, the ratio should be converted to a percentage and Rule 2 applied. This would apply, for example, to the case where the denominator is the number of individuals who smoke at all and the numerator is the number of individuals who smoke daily out of those who smoke at all.

Consider the case where the numerator is not a subset of the denominator, as for example, the ratio of the number of individuals who smoke daily or occasionally as compared to the number of individuals who do not smoke at all. The standard deviation of the ratio of the estimates is approximately equal to the square root of the sum of squares of each coefficient of variation considered separately multiplied by $\hat{R}$, where $\hat{R}$ is the ratio of the estimates ($\hat{R} = \hat{X}_1 / \hat{X}_2$). That is, the standard error of a ratio is:

$$\sigma_{\hat{R}} = \hat{R}\sqrt{\alpha_1^{\,2} + \alpha_2^{\,2}}$$

where α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively.

The coefficient of variation of $\hat{R}$ is given by $\sigma_{\hat{R}} / \hat{R} = \sqrt{\alpha_1^{\,2} + \alpha_2^{\,2}}$. The formula will tend to overstate the error, if $\hat{X}_1$ and $\hat{X}_2$ are positively correlated and understate the error if $\hat{X}_1$ and $\hat{X}_2$ are negatively correlated.

Rule 5: Estimates of differences of ratios

In this case, Rules 3 and 4 are combined. The CVs for the two ratios are first determined using Rule 4, and then the CV of their difference is found using Rule 3.

11.2 Examples of using the CV tables for categorical estimates

The following "real life" examples are included to assist users in applying the foregoing rules.

Example 1: Estimates of numbers possessing a characteristic (aggregates)

Suppose that a user estimates that 4,722,617 individuals smoke daily in Canada. How does the user determine the coefficient of variation of this estimate?

1) Refer to the CANADA level CV table.

2) The estimated aggregate (4,722,617) does not appear in the left-hand column (the "Numerator of Percentage" column), so it is necessary to use the smallest figure closest to it, namely 4,000,000.

3) The coefficient of variation for an estimated aggregate (expressed as a percentage) is found by referring to the first non-asterisk entry on that row, namely, 1.0%.

4) So the approximate coefficient of variation of the estimate is 1.0%. According to the Sampling Variability Guidelines presented in Section 10.4, the finding that there were 4,722,617 individuals who smoke daily is publishable with no qualifications.

Example 2 : Estimates of proportions or percentages possessing a characteristic

Suppose that the user estimates that 4,722,617/6,081,453=77.7% of individuals in Canada who smoke at all smoke daily. How does the user determine the coefficient of variation of this estimate?

1) Refer to the CANADA level CV table.

2) Because the estimate is a percentage which is based on a subset of the total population (i.e., individuals who smoke at all, that is to say, daily or occasionally), it is necessary to use both the percentage (77.7%) and the numerator portion of the percentage (4,722,617) in determining the coefficient of variation.

3) The numerator (4,722,617) does not appear in the left-hand column (the "Numerator of Percentage" column) so it is necessary to use the smallest figure closest to it, namely 4,000,000. Similarly, the percentage estimate does not appear as any of the column headings, so it is necessary to use the figure closest to it, 70.0%.

4) The figure at the intersection of the row and column used, namely 0.6% is the coefficient of variation (expressed as a percentage) to be used.

5) So the approximate coefficient of variation of the estimate is 0.6%. According to the Sampling Variability Guidelines presented in Section 10.4, the finding that 77.7% of individuals who smoke at all smoke daily can be published with no qualifications.

Example 3 : Estimates of differences between aggregates or percentages

Suppose that a user estimates that, among men, 2,535,367/13,078,499 = 19.4% smoke daily (estimate 1), while for women, this percentage is estimated at 2,187,250 / 13,476,931 = 16.2% (estimate 2). How does the user determine the coefficient of variation of the difference between these two estimates?

1) Using the CANADA level CV table in the same manner as described in example 2 gives the CV for estimate 1 as 1.5% (expressed as a percentage), and the CV for estimate 2 as 1.5% (expressed as a percentage).

2) Using rule 3, the standard error of a difference ($\hat{d} = \hat{X}_2 - \hat{X}_1$) is :

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

where $\hat{X}_1$ is estimate 1, $\hat{X}_2$ is estimate 2, and α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. The standard error of the difference $\hat{d} = (0.194 - 0.162) = 0.032$ is :

$$\sigma_{\hat{d}} = \sqrt{[0(.194)(0.015)]^2 + [(0.162)(0.015)]^2}$$
$$= 0.004$$

3) The coefficient of variation of $\hat{d}$ is given by $\sigma_{\hat{d}} / \hat{d} = 0.004/0.032 = 0.125$.

4) So the approximate coefficient of variation of the difference between the estimates is 12.5% (expressed as a percentage). According to the Sampling Variability Guidelines presented in Section 10.4, this estimate can be published with no qualifications.

Example 4 : Estimates of ratios

Suppose that the user estimates that 4,722,617 individuals smoke daily, while 1,358,836 individuals smoke occasionally. The user is interested in comparing the estimate of daily to occasional smokers in the form of a ratio. How does the user determine the coefficient of variation of this estimate?

1) First of all, this estimate is a ratio estimate, where the numerator of the estimate (= $\hat{X}_1$) is the number of individuals who smoke occasionally. The denominator of the estimate (= $\hat{X}_2$) is the number of individuals who smoke daily.

2) Refer to the CANADA level CV table.

3) The numerator of this ratio estimate is 1,358,836. The smallest figure closest to it is 1,000,000. The coefficient of variation for this estimate (expressed as a percentage) is found by referring to the first non-asterisk entry on that row, namely, 2.3%.

4) The denominator of this ratio estimate is 4,722,617. The figure closest to it is 4,000,000. The coefficient of variation for this estimate (expressed as a percentage) is found by referring to the first non-asterisk entry on that row, namely, 1.0%.

5) So the approximate coefficient of variation of the ratio estimate is given by rule 4, which is,

$$\alpha_{\hat{R}} = \sqrt{\alpha_1^{\ 2} + \alpha_2^{\ 2}},$$

That is,

$$\alpha_{\hat{R}} = \sqrt{(.023)^2 + (.01)^2}$$

$$= 0.025$$

where α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. The obtained ratio of occasional to daily smokers is 1,358,836/4,722,617 which is 0.29:1. The coefficient of variation of this estimate is 2.5% (expressed as a percentage), which is releasable with no qualifications, according to the Sampling Variability Guidelines presented in Section 10.4.

11.3 How to use the CV tables to obtain confidence limits

Although coefficients of variation are widely used, a more intuitively meaningful measure of sampling error is the confidence interval of an estimate. A confidence interval constitutes a statement on the level of confidence that the true value for the population lies within a specified range of values. For example a 95% confidence interval can be described as follows: if sampling of the population is repeated indefinitely, each sample leading to a new confidence interval for an estimate, then in 95% of the samples the interval will cover the true population value.

Using the standard error of an estimate, confidence intervals for estimates may be obtained under the assumption that under repeated sampling of the population, the various estimates obtained for a population characteristic are normally distributed about the true population value. Under this assumption, the chances are about 68 out of 100 that the difference between a sample estimate and the true population value would be less than one standard error, about 95 out of 100 that the difference would be less than two standard errors, and about 99 out of 100 that the differences would be less than three standard errors. These different degrees of confidence are referred to as the confidence levels.

Confidence intervals for an estimate, $\hat{X}$, are generally expressed as two numbers, one below the estimate and one above the estimate, as $\left(\hat{X} - k, \hat{X} + k\right)$, where k is determined depending upon the level of confidence desired and the sampling error of the estimate.

Confidence intervals for an estimate can be calculated directly from the Approximate Sampling Variability Tables by first determining from the appropriate table the coefficient of variation of the estimate $\hat{X}$, and then using the following formula to convert to a confidence interval CI:

$$CI_X = [\ \hat{X} - z\ \hat{X}\ \alpha_{\hat{X}}\ , \hat{X} + z\ \hat{X}\ \alpha_{\hat{X}}\ ]$$

where $\alpha_{\hat{X}}$ is determined coefficient of variation for $\hat{X}$, and

$z =$ 1 if a 68% confidence interval is desired
$z =$ 1.6 if a 90% confidence interval is desired
$z =$ 2 if a 95% confidence interval is desired
$z =$ 3 if a 99% confidence interval is desired.

Note: Release guidelines presented in section 10.4 which apply to the estimate also apply to the confidence interval. For example, if the estimate is not releasable, then the confidence interval is not releasable either.

11.4 Example of using the CV tables to obtain confidence limits

A 95% confidence interval for the estimated proportion of individuals who smoke daily from those who smoke at all (from example 2, sub-section 11.2) would be calculated as follows:

$\hat{X}$ = 0.777

z = 2

$\alpha_{\hat{X}}$ = 0.006 is the coefficient of variation of this estimate as determined from the tables.

$CI_{\hat{x}}$ = {0.777 - (2) (0.777) (0.006) , 0.777 + (2) (0.777) (0.006)}

$CI_{\hat{x}}$ = {0.768 , 0.786}

11.5 How to use the CV tables to do a Z-test

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The sample estimates can be numbers, averages, percentages, ratios, etc. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical.

Let $\hat{X}_1$ and $\hat{X}_2$ be sample estimates for 2 characteristics of interest. Let the standard error on the difference $\hat{X}_1 - \hat{X}_2$ be $\sigma_{\hat{d}}$. If the ratio of $\hat{X}_1 - \hat{X}_2$ over $\sigma_{\hat{d}}$ is between -2 and 2, then no conclusion about the difference between the characteristics is justified at the 5% level of significance. If however, this ratio is smaller than -2 or larger than +2, the observed difference is significant at the 0.05 level.

11.6 Example of using the CV tables to do a Z-test

Let us suppose we wish to test, at 5% level of significance, the hypothesis that there is no difference between the proportion of men who smoke daily AND the proportion of women who smoke daily. From example 3, sub-section 11.2, the standard error of the difference between these two estimates was found to be = 0.004. Hence,

$$z = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}} = \frac{0.194 - 0.162}{0.004} = \frac{0.032}{0.004} = 8$$

Since $z = 8$ is greater than 2, it must be concluded that there is a significant difference between the two estimates at the 0.05 level of significance. Note that the two sub-groups compared are considered as being independent, so the test is correct.

11.7 Exact variances/coefficients of variation

All coefficients of variation in the Approximate Sampling Variability Tables (CV Tables) are indeed approximate and, therefore, unofficial.

The computation of exact coefficients of variation is not a straightforward task since there is no simple mathematical formula that would account for all CCHS sampling frame and weighting aspects. Therefore, other methods such as resampling methods must be used in order to estimate measures of precision. Among these methods, the bootstrap method is the one recommended for analysis of CCHS data.

The computation of coefficients of variation (or any other measure of precision) with the use of the bootstrap method requires access to information that is considered confidential and not available on the PUMF. This computation must be done using the Master file. Access to the Master file is discussed in section 12.3.

For the computation of coefficients of variation, the bootstrap method is advised. A macro program, called “Bootvar”, was developed in order to give users easy access to the bootstrap method. The Bootvar program is available in SAS and SPSS formats, and is made up of macros that calculate the variances of totals, ratios, differences between ratios, and linear and logistic regressions.

There are a number of reasons why a user may require an exact variance. A few are given below.

Firstly, if a user desires estimates at a geographic level other than those available in the tables (for example, at the rural/urban level), then the CV tables provided are not adequate. Coefficients of variation of these estimates may be obtained using "domain" estimation techniques through the exact variance program.

Secondly, should a user require more sophisticated analyses such as estimates of parameters from linear regressions or logistic regressions, the CV tables will not provide correct associated coefficients of variation. Although some standard statistical packages allow sampling weights to be incorporated in the analyses, the variances that are produced often do not take into account the stratified and clustered nature of the design properly, whereas the exact variance program would do so.

Thirdly, for estimates of quantitative variables, separate tables are required to determine their sampling error. Since most of the variables for the CCHS are primarily categorical in nature, this has not been done. Thus, users wishing to obtain coefficients of variation for quantitative variables can do so through the exact variance program. As a general rule, however, the coefficient of variation of a quantitative total will be larger than the coefficient of variation of the corresponding category estimate (i.e., the estimate of the number of persons contributing to the quantitative estimate). If the corresponding category estimate is not releasable, the quantitative estimate will not be either. For example, the coefficient of variation of the estimate of the total number of cigarettes smoked each day by individuals who smoke daily would be greater than the coefficient of variation of the corresponding estimate of the number of individuals who smoke daily. Hence if the coefficient of variation of the latter is not releasable, then the coefficient of variation of the corresponding quantitative estimate will also not be releasable.

Lastly, should users find themselves in a position where they can use the CV tables, but this renders a coefficient of variation in the "marginal" range (16.6% - 33.3%), the user should release the associated estimate with a warning cautioning users of the high sampling variability associated with the estimate. This would be a good opportunity to recalculate the coefficient of variation through the exact variance program to find out if it is releasable without a qualifying note. The reason for this is that the coefficients of variation produced by the tables are based on a wide range of variables and are therefore considered crude, whereas the exact variance program would give an exact coefficient of variation associated with the variable in question.

11.8 Release cut-offs for the CCHS

Appendix E presents tables giving the minimum cut-offs for estimates of totals at the Canada, provincial, health region and CLSC levels and those for various age groups at the Canada level. Estimates smaller than the value given in the "Marginal" column may not be released under any circumstances.

12. File usage

This section begins by describing the ***weight variable*** of the PUMF and explains how it should be used when doing tabulations. This is followed by an explanation of the variable naming convention that is employed by the CCHS. The last part of the section discusses alternate approaches for data access available to analysts.

12.1 Use of weight variable

The weight variable **WTSE_M** represents the CCHS Cycle 3.1 sampling weight. For a given respondent, the sampling weight can be interpreted as the number of people the respondent represents in the population. This weight must always be used when computing statistical estimates in order to make inference at the population level possible. The production of unweighted estimates is not recommended. The sample allocation, as well as the survey design specifics can cause such results to not correctly represent the population. Refer to section 8 on weighting for a more detailed explanation on the creation of this weight.

12.2 Variable naming convention

The CCHS adopted a variable naming convention that allows data users to easily use and identify the data based on module and cycle. The variable naming convention includes the following mandatory requirements: restrict variable names to a maximum of 8 characters for ease of use by analytical software products; identify the survey occasion (Cycle 2.1, 1.2 …) in the name; and allow conceptually identical variables to be easily identifiable over survey occasions. The variable names for these identical modules and questions should only differ in the cycle position identifying the particular survey occasion in which they were collected.

12.2.1 Variable name component structure in CCHS

Each of the eight characters in a variable name contains information about the type of data contained in the variable.

Positions 1-3: Module/Questionnaire section name
Position 4: Survey cycle
Position 5: Variable type
Positions 6-8: Question number

For example: The variable from question 202, Smoking Module, CCHS Cycle 3.1 (SMKE_202):

Position 1-3:	SMK	depression module
Position 4:	E	Cycle 3.1
Position 5:	_	underscore (_ = collected data)
Position 6-8:	202	question number & answer option

12.2.2 Positions 1-3: variable / questionnaire section name

The following values are used for the section name component of the variable name:

ACC	Access to health care services	MAM	Mammography
ADM	Administration	MED	Medication use
ALC	Alcohol use	MEX	Maternal experiences
BPC	Blood pressure check	NDE	Nicotine dependence
BRX	Breast examinations	OH2	Oral health 2
BSX	Breast self examinations	ORG	Voluntary organizations
CCC	Chronic conditions	PAC	Physical activities
CCS	Colorectal cancer exams	PAP	PAP smear test
CIH	Changes made to improve health	PAS	Patient satisfaction
CMH	Contacts with mental health professionals	PCU	Physical check-up
CPG	Canadian Problem Gambling Index	PSA	Prostate cancer screening
CST	Childhood and adult stressors	RAC	Restriction of activities
DEN	Dental visits	REP	Injuries (repetitive strain)
DHH	Demographics and household	SAC	Sedentary activities
DIA	Diabetes Care	SAM	Sample Identifiers
DIS	Distress	SCA	Smoking cessation aids
DPS	Depression	SCH	Smoking - stages of change
EDU	Education	SDC	Socio-demographic characteristics
ETS	Exposure to second-hand smoke	SFE	Self-esteem
EYX	Eye Examinations	SFR	Health status – SF-36
FDC	Food choices	SMK	Smoking
FLU	Flu shots	SLP	Sleep
FSC	Food security	SPC	Smoking – physician counselling
FVC	Fruit and vegetable consumption	SSA	Social support – availability
GEN	General health	SSB	Sun safety
GEO	Geographic identifiers	SSU	Social support – utilization
HCS	Health care system satisfaction	STC	Stress – coping
HCU	Health care utilization	STS	Stress – sources
HMC	Home care	SUI	Suicidal thoughts and attempts
HUI	Health Utility Index (HUI)	SWL	Satisfaction with life
HWT	Height and weight	SXB	Sexual behaviour
IDG	Illicit drugs	TWD	Two-week disability
INC	Income	UPE	Use of protective equipment
INJ	Injuries	WST	Work stress
INS	Insurance coverage	WTM	Waiting times
LBF	Labour force	WTS	Sample weights
		YSM	Youth smoking

12.2.3 Position 4: cycle

Cycle **Description**

A Cycle 1.1: Canadian Community Health Survey
- Regional level survey, stratified by health region
- Common content and optional content selected by health region
- Estimates for health regions, provinces, territories and Canada

B Cycle 1.2: Canadian Community Health Survey, Mental Health and Well-Being
- Provincial level survey
- Focus content with additional, general content
- Estimates for the provinces, territories and Canada

C Cycle 2.1: Canadian Community Health Survey
- Regional level survey, stratified by health region
- Common content and optional content selected by health region
- Estimates for health regions, provinces, territories and Canada

D Cycle 2.2: Canadian Community Health Survey, Nutrition
- Provincial level survey
- Focus content with additional, general content
- Estimates for the provinces, territories and Canada

E Cycle 3.1: Canadian Community Health Survey
- Regional level survey, stratified by health region
- Common content and optional content selected by health region
- Estimates for health regions, provinces, territories and Canada

12.2.4 Position 5: variable type

_	Collected variable	A variable that appeared directly on the questionnaire
C	Coded variable	A variable coded from one or more collected variables (e.g., SIC, Standard Industrial Classification code)
D	Derived variable	A variable calculated from one or more collected or coded variables, usually calculated during head office processing (e.g., Health Utility Index)
F	Flag variable	A variable calculated from one or more collected variables (like a derived variable), but usually calculated by the data collection computer application for later use during the interview (e.g., work flag)

G	Grouped variable	Collected, coded, suppressed or derived variables collapsed into groups (e.g., age groups)

12.2.5 Positions 6-8: variable name

In general, the last three positions follow the variable numbering used on the questionnaire. The letter "Q" used to represent the word "question" is removed, and all question numbers are presented in a two- digit format. For example, question Q01A in the questionnaire becomes simply 01A, and question Q15 becomes simply 15.

For questions which have more than one response option, the final position in the variable naming sequence is represented by a letter. For this type of question, new variables were created to differentiate between a “yes” or “no” answer for each response option. For example, if Q2 had 4 response options, the new questions would be named Q2A for option 1, Q2B for option 2, Q2C for option 3, etc. If only options 2 and 3 were selected, then Q2A = No, Q2B = Yes, Q2C = Yes and Q2D = No.

12.3 Access to master file data

In order to protect the confidentiality of respondents participating in the survey, the PUMFs must meet stringent security and confidentiality standards required by the Statistics Act before they are released for public access. To ensure that these standards have been achieved, each PUMF goes through a formal review process to ensure that an individual cannot be identified. Rare values in variables that may lead to identification of an individual are suppressed on the file or are collapsed to broader categories so that individual disclosure is minimized. Frequently, these are the variables that are most critical for doing a complete and comprehensive analysis of the survey data. Since a significant amount of resources is spent on collecting these data, ensuring that the PUMFs reach their full analytical potential is important for a complete return on the statistical investment.

One approach for any user is the production of custom tabulations done by the Client Custom Services staff in Health Statistics Division. This service allows users who do not possess knowledge of tabulation software products to get custom results. The results are screened for confidentiality and reliability concerns before release. There is a charge for this service.

A second approach is the Research Data Centres Program, which allows researchers to submit to Statistics Canada, a research project that uses data from the Master File. These projects are accepted based on a set of specific rules. When the project is accepted, the researcher is designated as a "deemed employee" of Statistics Canada for the duration of the research, and given access to the Master File data from designated Statistics Canada sites. For more information, please consult the Statistics Canada webpage http://www.statcan.ca/english/rdc/index.htm.

Finally, the remote access service to the survey master file is another way to have access to these data if for some reason, the user cannot access a RDC. Each purchaser of the microdata product can be supplied with a ‘dummy’ test master file and a corresponding record layout. With this, the user can spend time developing a set of analytical computer programs using the test file to confirm

that the program commands are functioning correctly. At that point, the code for the custom tabulations is then sent via e-mail to *cchs-escc@statcan.ca*. The code will then be transferred into Statistics Canada's internal secured network and processed using the appropriate master file of CCHS Cycle 3.1 data. Estimates generated will be released to the user, subject to meeting the guidelines for analysis and release outlined in Section 10 of this document. Results are screened for confidentiality and reliability concerns and, once these have been addressed, the output is returned to the client. There is no charge for this service.

APPENDIX A

CANADIAN COMMUNITY HEALTH SURVEY (CCHS)
CYCLE 3.1 (2005)

MASTER FILE

GUIDELINES FOR THE USE OF SUB-SAMPLE VARIABLES

Statistics Canada

June 2006

CCHS Cycle 3.1 sub-samples

The CCHS Cycle 3.1 questionnaire consists of three types of questionnaire modules:

1. **Common content -** Questionnaire modules asked of all respondents in all health regions.
2. **Optional content -** Questionnaire modules included in the questionnaire at the discretion of health regions. When a module was included as optional content in a given health region, it was asked of all respondents in the region. Although the aim of optional content was to permit calculation of health region estimates, it should be noted that in Cycle 3.1 all health regions within each of the 10 provinces selected the same optional content, which yields the possibility of calculating provincial estimates.
3. **Sub-sample Content -** Questionnaire modules which were asked only of a subset of respondents. The aim was to permit calculation of provincial and national estimates while minimising response burden. Three sets of sub-sample modules were asked to three (separate but overlapping) sub-samples:

   Sub-Sample 1: Health Utility Index (HUI), Fruit and Vegetable Consumption and Labour Force (long form).

   Sub-Sample 2: Measured Height and Weight

   Sub-Sample 3: Access to Health Care Services, Waiting Times and Patient Satisfaction

Note: Sub-sample 3 completely replaces the Health Services Access Survey (HSAS). In 2000/01, HSAS was conducted as a followup to CCHS. In 2003, HSAS was partially integrated with CCHS.

Variable naming convention for sub-sample modules

Variables from common and optional content modules are designated with a "E" in position 4 of the variable name; eg. SMK**E**_201.

Variables from sub-sample content modules are designated with a "Z" in position 4 of the variable name; eg. ACC**Z**_01.

With the exception of Measured Height and Weight, each of the sub-sample modules was also made available to health regions as optional content. For each of these modules, there are two sets of variables; eg. HUIE_28 and HUI**Z**_28. These variables are applicable or not applicable for a given respondent according to whether or not the respondent is a member of the relevant sub-sample **and** whether or not the respondent lives in a health region where the module was chosen as optional content. Table 1 illustrates the possibilities:

Table 1. Possible not-applicable values for typical sub-sample variable

Member of sub-sample 1	Respondent lives in HR where Health Utility Index (HUI) was chosen as optional content	
	YES	**NO**
YES	HUIE_28 and HUIZ_28 are applicable	HUIE_28 is not applicable and HUIZ_28 is applicable
NO	HUIE_28 is applicable and HUIZ_28 is not applicable	HUIE_28 and HUIZ_28 are not applicable

Sub-sample content appears on physically separate files

To encourage the appropriate use of CCHS data, sub-sample content is provided on physically separate files. Each file has a corresponding sample weight and set of bootstrap weights which must be used to produce valid estimates for all variables on the file. Table 2 presents a description of the set of four data files released in Cycle 3.1.

Table 2. File names and content of CCHS 3.1 data files

File	File name	Sampling weight	Bootstrap weights file	Variables included
Main Master file	HS.txt	WTSE_M	b5.txt	All common and all optional modules. For modules which are part of both optional and sub-sample content, only the "E" set of variables is included.
Sub-sample 1 master file	HS_S1.txt	WTSE_S1M	b5_s1.txt	All common modules plus the "**Z**" set of variables for the Health Utility Index, Fruit and Vegetable Consumption and Labour Force (long form) modules.
Sub-sample 2 master file	HS_S2.txt	WTSE_S2M	b5_s2.txt	All common modules plus the "**Z**" set of variables for the Measured Height/Weight module.
Sub-sample 3 master file	HS_S3.txt	WTSE_S3M	b5_s3.txt	All common modules, plus the "**Z**" set of variables for the Access to Health Care Services, Waiting Times and Patient Satisfaction modules.

Analysis combining sub-sample and optional content

The aim of the CCHS sub-sample modules is to permit the calculation of estimates at the national level. Provincial/territorial level estimates are also available for all sub-sample modules except Measured Height and Weight. The sub-samples, and their associated weights, are not intended to support health region level estimates.

In Cycle 3.1, some sub-sample modules were also selected by all health regions of at least one province as optional content, as shown in Table 3.

Table 3 Sub-sample modules which were also selected as optional content

Sub-Sample	Provinces where sub-sample modules were chosen by all health regions in a province
1	• HUI (British Columbia) • Fruit and Vegetable Consumption (PEI, Ontario, Alberta, B.C.)
3	• Access to Health Care Services, Waiting Times (New Brunswick)

When a module from **sub-sample 1 or sub-sample 3** is chosen as optional content by all health regions in a province it is possible to calculate health region and provincial estimates of the variables in that module using the main **master file and the master sampling weight**. This offers the advantage of a larger sample size therefore smaller sampling error. (Sample sizes for the main master file and each of the sub-sample files are summarised in Tables 4 through 7). The provincial estimate can then be compared to the estimate for other provinces calculated using the appropriate sub-sample file and sub-sample sampling weight.

Note: It is not possible to produce valid estimates from cross-tabulations or multivariate model using a combination of variables from different sub-samples. The sub-samples have distinct sample sizes and thus, different weights.

Table 4. Sample sizes by province/territory for main master file

Province / Territory	Sample size of main master file
CANADA	**132,947**
Newfoundland and Labrador	4,111
Prince Edward Island	2,031
Nova Scotia	5,066
New Brunswick	5,100
Quebec	29,891
Ontario	41,766
Manitoba	7,352
Saskatchewan	7,765
Alberta	11,800
British Columbia	15,407
Yukon	868
Northwest Territories	1,007
Nunavut	783

Table 5. Sample sizes by province/territory for sub-sample 1
(Health Utility Index [HUI], Fruit and Vegetable Consumption and Labour Force [long form])

Province / Territory	Sample size of sub-sample 1
CANADA	**32,153**
Newfoundland and Labrador	1,571
Prince Edward Island	1,008
Nova Scotia	1,960
New Brunswick	1,736
Quebec	5,446
Ontario	6,622
Manitoba	2,104
Saskatchewan	2,075
Alberta	3,225
British Columbia	3,748
Yukon	868
Northwest Territories	1,007
Nunavut	783

1. Fifty-five respondents who refused to answer the first question in each module (HUIZ_01, FVCZ_1A and LFSZ_01) were excluded from sub-sample 1.

Table 6. Sample sizes by province/territory for sub-sample 2
(Measured Height and Weight)

Province / Territory	Sample size of sub-sample 2[1]
CANADA	**4,735**
Newfoundland and Labrador	85
Prince Edward Island	17
Nova Scotia	127
New Brunswick	93
Quebec	1,185
Ontario	1,785
Manitoba	186
Saskatchewan	172
Alberta	458
British Columbia	627
Yukon	0
Northwest Territories	0
Nunavut	0

1. Includes pregnant women

Table 7. Sample sizes by province/territory for sub-sample 3
(Access to Health Care Services, Waiting Times and Patient Satisfaction)

Province / Territory	Sample size of sub-sample 3[1]
CANADA	**35,968**
Newfoundland and Labrador	2,798
Prince Edward Island	1,882
Nova Scotia	3,035
New Brunswick	3,039
Quebec	4,558
Ontario	4,692
Manitoba	3,076
Saskatchewan	3,199
Alberta	3,458
British Columbia	3,802
Yukon	804
Northwest Territories	925
Nunavut	700

1. Sub-sample 3 includes only those aged 15 and older.

Other considerations

Data dictionary

Separate data dictionary reports, including universe statements and frequencies, are provided for the main master file and each of the three sub-sample files.

In the master file data dictionary reports, optional content modules are treated in the same way as for CCHS 1.1 and CCHS 2.1. For each module, a flag indicates whether a given respondent lives in a health region where the module was selected as optional content. When the flag is equal to 2 (No), all variables in the module have not applicable values. For example, the variable PASEFOPT indicates whether the module Patient Satisfaction is applicable for a given respondent.

Population totals

The sampling weights for sub-samples 1 and 2 are calibrated so that they add to total Canadian population aged 12 and older (27,131,965). The sampling weights for sub-sample 3 (in which all modules apply to respondents aged 15 and older) are calibrated so that they add to the total Canadian population aged 15 and older (25,877,399).

Differences in calculation of common content variables using different files

Variables from common content modules can be estimated using any of the four files provided. Depending on which file is used, very small differences will be observed.

All official Statistics Canada estimates of variables from common modules are based on the master file sampling weight.

Questions

For any questions related to the appropriate use of CCHS sub-sample data, please contact the Data Access Unit at:

1 (613) 951–1653
cchs-escc@statcan.ca

APPENDIX B: Health regions

Health regions for dissemination, before PUMF		Health regions for PUMF	
Newfoundland and Labrador		**Newfoundland and Labrador**	
1011	Eastern Health Authority	1011	Eastern Health Authority
1012	Health and Community Services Central Region	1012	Health and Community Services Central Region
1013	Health and Community Services Western Region	1013	Western Region/Labrador-Grenfell
1014	Labrador-Grenfell Health Authority		
Prince Edward Island		**Prince Edward Island**	
1101	West Prince Health Region	1101	West Prince/East Prince/Kings
1102	East Prince Health Region		
1104	Kings Health Region		
1103	Queens Health Region	1103	Queens Health Region
Nova Scotia		**Nova Scotia**	
1201	Zone 1	1201	Zone 1
1202	Zone 2	1202	Zone 2
1203	Zone 3	1203	Zone 3
1204	Zone 4	1204	Zone 4
1205	Zone 5	1205	Zone 5
1206	Zone 6	1206	Zone 6
New Brunswick		**New Brunswick**	
1301	Region 1	1301	Region 1
1302	Region 2	1302	Region 2
1303	Region 3	1303	Region 3
1304	Region 4	1304	Region 4/5
1305	Region 5		
1306	Region 6	1306	Region 6/7
1307	Region 7		
Quebec		**Quebec**	
2401	Région du Bas-Saint-Laurent	2401	Région du Bas-Saint-Laurent
2402	Région du Saguenay - Lac-Saint-Jean	2402	Région du Saguenay - Lac-Saint-Jean
2403	Région de Québec	2403	Région de Québec
2404	Région de la Mauricie et du Centre-du-Québec	2404	Région de la Mauricie et du Centre-du-Québec
2405	Région de l'Estrie	2405	Région de l'Estrie
2406	Région de Montréal-Centre	2406	Région de Montréal-Centre
2407	Région de l'Outaouais	2407	Région de l'Outaouais
2408	Région de l'Abitibi-Témiscamingue	2408	Région de l'Abitibi-Témiscamingue
2409	Région de la Côte-Nord	2409	Région de la Côte-Nord
2410	Région du Nord-du-Québec		
2411	Région de la Gaspésie - Îles-de-la-Madeleine	2411	Région de la Gaspésie - Îles-de-la-Madeleine
2412	Région de la Chaudière-Appalaches	2412	Région de la Chaudière-Appalaches
2413	Région de Laval	2413	Région de Laval
2414	Région de Lanaudière	2414	Région de Lanaudière
2415	Région des Laurentides	2415	Région des Laurentides
2416	Région de la Montérégie	2416	Région de la Montérégie

Health regions for dissemination, before PUMF		Health regions for PUMF	
Ontario		**Ontario**	
3526	Algoma	3526	Algoma
3527	Brant	3527	Brant
3530	Durham	3530	Durham
3531	Elgin-St. Thomas	3531	Elgin-St. Thomas
3533	Grey Bruce	3533	Grey Bruce
3534	Haldimand-Norfolk	3534	Haldimand-Norfolk
3535	Haliburton, Kawartha, Pine Ridge	3535	Haliburton, Kawartha, Pine Ridge
3536	Halton	3536	Halton
3537	Hamilton	3537	Hamilton
3538	Hastings and Prince Edward	3538	Hastings and Prince Edward
3539	Huron	3539	Huron/Perth
3554	Perth		
3540	Chatham-Kent	3540	Chatham-Kent
3541	Kingston, Frontenac and Lennox and Addington	3541	Kingston, Frontenac and Lennox and Addington
3542	Lambton	3542	Lambton
3543	Leeds, Grenville and Lanark	3543	Leeds, Grenville and Lanark
3544	Middlesex-London	3544	Middlesex-London
3546	Niagara	3546	Niagara
3547	North Bay Parry Sound	3547	North Bay Parry Sound/Timiskaming
3563	Timiskaming		
3549	Northwestern	3549	Northwestern
3551	Ottawa	3551	Ottawa
3552	Oxford	3552	Oxford
3553	Peel	3553	Peel
3555	Peterborough	3555	Peterborough
3556	Porcupine	3556	Porcupine
3557	Renfrew	3557	Renfrew
3558	Eastern Ontario	3558	Eastern Ontario
3560	Simcoe Muskoka	3560	Simcoe Muskoka
3561	Sudbury	3561	Sudbury
3562	Thunder Bay	3562	Thunder Bay
3565	Waterloo	3565	Waterloo
3566	Wellington-Dufferin-Guelph	3566	Wellington-Dufferin-Guelph
3568	Windsor-Essex	3568	Windsor-Essex
3570	York	3570	York
3595	City of Toronto	3595	City of Toronto
Manitoba		**Manitoba**	
4610	Winnipeg RHA	4610	Winnipeg RHA
4615	Brandon RHA	4615	Brandon/Assiniboine
4645	Assiniboine RHA		
4620	North Eastman RHA	4620	North Eastman/South Eastman
4625	South Eastman RHA		
4630	Interlake RHA	4630	Interlake RHA
4640	Central RHA	4640	Central RHA
4660	Parkland RHA	4660	Parkland/Norman/Burntwood-Churchill
4670	Norman RHA		
4685	Burntwood RHA/Churchill RHA		

Health regions for dissemination, before PUMF		Health regions for PUMF	
Saskatchewan		**Saskatchewan**	
4701	Sun Country RHA	4701	Sun Country/Five Hills/Cypress
4702	Five Hills RHA		
4703	Cypress RHA		
4704	Regina Qu'Appelle RHA	4704	Regina Qu'Appelle RHA
4705	Sunrise RHA	4705	Sunrise/Kelsey Trail
4708	Kelsey trail RHA		
4706	Saskatoon RHA	4706	Saskatoon RHA
4707	Heartland RHA	4707	Heartland/Prairie North
4710	Prairie North RHA		
4709	Prince Albert Parkland RHA	4709	Prince Albert Parkland/Mamawetan-Keewatin Yatthé-Athabaska
4714	Mamawetan Churchill River RHA/Keewatin Yatthé RHA/Athabaska HA		
Alberta		**Alberta**	
4820	Chinook Regional Health Authority	4820	Chinook Regional Health Authority
4821	Palliser Health Region	4821	Palliser Health Region
4822	Calgary Health Region	4822	Calgary Health Region
4823	David Thompson Regional Health Authority	4823	David Thompson Regional Health Authority
4824	East Central Health	4824	East Central Health
4825	Capital Health	4825	Capital Health
4826	Aspen Regional Health Authority	4826	Aspen Regional Health Authority
4827	Peace Country Health	4827	Peace Country/Northern Lights
4828	Northern Lights Health Region		
British Columbia		**British Columbia**	
5911	East Kootenay	5911	East Kootenay
5912	Kootenay-Boundary	5912	Kootenay-Boundary
5913	Okanagan	5913	Okanagan
5914	Thompson/Cariboo	5914	Thompson/Cariboo
5921	Fraser East	5921	Fraser East
5922	Fraser North	5922	Fraser North
5923	Fraser South	5923	Fraser South
5931	Richmond	5931	Richmond
5932	Vancouver	5932	Vancouver
5933	North Shore/Coast Garibaldi	5933	North Shore/Coast Garibaldi
5941	South Vancouver Island	5941	South Vancouver Island
5942	Central Vancouver Island	5942	Central Vancouver Island
5943	North Vancouver Island	5943	North Vancouver Island
5951	Northwest	5951	Northwest/Northeast
5953	Northeast		
5952	Northern Interior	5952	Northern Interior
Territories		**Territories**	
6001	Yukon Territory	6001	Yukon Territory/Northwest Territories/Nunavut
6101	Northwest Territories		
6201	Nunavut		

Quebec buy-in health regions before PUMF		Quebec buy-in health regions for PUMF	
Quebec buy-ins			
2401	**Région du Bas-Saint-Laurent**	2401	**Région du Bas-Saint-Laurent**
1	La Matapédia	1	Est
2	Matane		
3	La Mitis		
4	Rimouski-Neigette		
5	Les Basques	5	Ouest
6	Rivière-du-Loup		
7	Témiscouata		
8	Kamouraska		
2406	**Région de Montréal-Centre**	2406	**Région de Montréal-Centre**
1	Pierrefonds et Lac Saint-Louis	1	Pierrefonds et Lac Saint-Louis
2	LaSalle et Vieux Lachine	2	LaSalle et Vieux Lachine
3	Verdun/Côte Saint-Paul,Saint-Henri,Pointe Saint-Charles	3	Verdun/Côte Saint-Paul,Saint-Henri,Pointe Saint-Charles
4	René-Cassin et NDG/Montréal-Ouest	4	René-Cassin et NDG/Montréal-Ouest
5	Côte-des-Neiges,Métro et Parc Extension	5	Côte-des-Neiges,Métro et Parc Extension
6	Nord de l'Île et Saint-Laurent	6	Nord de l'Île et Saint-Laurent
7	Ahuntsic et Montréal-Nord	7	Ahuntsic et Montréal-Nord
8	La Petite Patrie et Villeray	8	La Petite Patrie et Villeray
9	Faubourgs, Plateau Mont-Royal et Saint-Louis du Parc	9	Faubourgs, Plateau Mont-Royal et Saint-Louis du Parc
10	Saint-Michel et Saint-Léonard	10	Saint-Michel et Saint-Léonard
11	Hochelaga-Maisonneuve, Olivier-Guimond et Rosemont	11	Hochelaga-Maisonneuve, Olivier-Guimond et Rosemont
12	Rivière-des-Prairies, Mercier-Est/Anjou et Pointe-aux-Trembles/Montréal-Est	12	Rivière-des-Prairies, Mercier-Est/Anjou et Pointe-aux-Trembles/Montréal-Est
2413	**Région de Laval**	2413	**Région de Laval**
1	Est	1	Est
2	Ouest	2	Ouest