Statistics Canada
Health Statistics Division

Canadian Community Health Survey

2003

User Guide for the Public Use Microdata File

Published by authority of the Minister responsible for Statistics Canada

January 2005

Catalogue no. 82M0013GPE

Ottawa

La version française de cette publication est disponible (no 82M0013GPF au catalogue)

This documentation can also be found in CCHS 2003 CD-ROM
(Catalogue no. 82M0013XCB)

Note of Appreciation

Canada owes the success of its statistical system to a long-standing cooperation between Statistics Canada, the citizens of Canada, its businesses and governments. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Health Statistics Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6 (telephone: (613) 951-1746).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our Web site.

National inquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Depository Services Program inquiries	**1 800 700-1033**
Fax line for Depository Services Program	**1 800 889-9734**
E-mail inquiries	**infostats@statcan.ca**
Web site	**www.statcan.ca**

Ordering and subscription information

This product, Catalogue no. 82M0021GPE, is published as a standard printed publication at a price of CAN $50.00. The following additional shipping charges apply for delivery outside Canada:

	Single issue
United States	CAN $6.00
Other countries	CAN $10.00

All prices exclude sales taxes.

The printed version of this publication can be ordered by

- Phone (Canada and United States) **1 800 267-6677**
- Fax (Canada and United States) **1 877 287-4369**
- E-mail **infostats@statcan.ca**
- Mail Statistics Canada
  Dissemination Division
  Circulation Management
  120 Parkdale Avenue
  Ottawa, Ontario K1A 0T6
- And, in person at the Statistics Canada Reference Centre nearest you, or from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136.

CANADIAN COMMUNITY HEALTH SURVEY (CCHS) CYCLE 2.1 (2003)

PUBLIC USE MICRODATA FILE DOCUMENTATION

STATISTICS CANADA

JANUARY 2005

TABLE OF CONTENTS

List of Appendices

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" is a smaller survey designed to provide provincial level results on specific focused health topics.

This Microdata File contains data collected in the third year of collection for the CCHS (Cycle 2.1). Information was collected between January 2003 and December 2003, for 126 health regions, covering all provinces and territories. The CCHS (Cycle 2.1) collects responses from persons aged 12 or older, living in private occupied dwellings. Excluded from the sampling frame are individuals living on Indian Reserves and on Crown Lands, institutional residents, full-time members of the Canadian Armed 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 2.1) microdata file, 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 the needs, the Canadian Institute for Health Information (CIHI), Statistics Canada and Health Canada have joined forces to create a Health Information Roadmap.

The Roadmap is a direct response to the concerns and desires of more than 500 individuals representing a wide range of organizations and interest groups. Early in 1999, the three national organizations listed above conducted a broadly based national consultation on health information needs. Participants stressed that national agencies must work together to strengthen Canada's health information system, and must build on and contribute to the considerable investments and expertise at local, regional, and provincial/territorial levels.[2]

The Roadmap represents an important contribution to building a comprehensive national health information system and infrastructure to provide Canadians with the information they need to maintain and improve Canada's health system and the population's health.[3] What is needed is a co-ordinated plan of action. No single government or organization can combat the above–noted problems alone. Co-operation at all levels – national, provincial, territorial, regional and local health organizations – is a prerequisite for success.[4]

The plan of action starts by seeking answers to two crucial questions[5]:

1. How healthy is the health care system?
2. How healthy are Canadians?

The first question encompasses the effectiveness, efficiency and responsiveness of the health care system. Generally, an effective, efficient and responsive health care system is one that offers the quality of care Canadians expect.[6]

The second question is broader, and addresses the basic objective of the system: Is the health of Canadians improving? To answer this, a strong health information system is needed.[7] This information system must embrace six principle characteristics[8].

[1] 1999. Health Information Roadmap Responding to Needs, Health Canada, Statistics Canada. p.3.
[2] 1999. Ibid. p.1.
[3] 1999. Ibid. p.1.
[4] 1999. Ibid. p.3.
[5] 1999. Ibid. p.3.
[6] 1999. Ibid. p.3.
[7] 1999. Ibid. p.5.
[8] Expansion on these characteristics is described in Health Information Roadmap: Responding to Needs, 1999, Canadian Institute for Health Information. ISBN 1-895581-30-3. (http://www.cihi.ca)

The information system must be:

- secure and respectful of Canadians' privacy;
- consistent;
- relevant;
- integrable;
- flexible;
- user-friendly and accessible.

This health information system needs to be timely, provide person-oriented information, and have common data standards with other Canadian health surveys, such as the National Population Health Survey (NPHS). The new system must also provide: new or expanded data sets; data on health services; data on outcomes, health status and non-medical determinants of health; data on outcomes of selected health interventions; implement special studies involving priority issues; data on costs per service; information exchange protocols; expanded analytical and dissemination capacity, and public reports on the health care system.[9]

Given 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[10].

The purpose of this publication, the Public Use Microdata File, 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.

[9] 1999. Ibid. p.11-14.
[10] 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 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;

- Create a flexible survey instrument that :
  - meets specific health region data gaps;
  - develops focused survey content for key data; and
  - deals with emerging health and health care issues as they arise.

As a key component of the Population Health Survey Program of Statistics Canada, the CCHS helps 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 this study. The second sub-section describes the common content in detail followed by a sub-section explaining the optional content of the CCHS (Cycle 2.1).

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.

Whereas consultations prior to CCHS Cycle 1.1 used a combination of qualitative and quantitative methods to identify the relative priority of broad topic areas, the primary objective of the Cycle 2.1 consultations was to identify new and emerging issues for which a data gap existed.

Based on these consultations, a list of topics to be included in Cycle 2.1 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 2.1 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 to Physical Activities and Two-week Disability. Table 4.1 outlines the common content for the CCHS (Cycle 2.1) as identified in the consultations. These common content topics, transformed into survey questions, were asked of all respondents in all health regions.

Table 4.1 CCHS Cycle 2.1 Common Content Modules

• Alcohol • Chronic conditions • Changes made to improve health • Exposure to second-hand smoke • Food insecurity • Flu Shots • Fruit and vegetable consumption • General health • Health care utilization • Height / weight • Home care • Injuries • Mammography • Maternal experiences • Oral health(1) – common • PAP smear test	• Physical activities • Repetitive strain • Restriction of activities • Sexual behaviour • Smoking • Two-week disability • Voluntary organizations • Youth smoking • Education • Geographic identifiers • Household composition • Income • Insurance coverage • Labour force • Socio-demographic characteristics

4.3 Optional Content

The topic content of the optional modules also emerged from the consultation process (see Table 4.2). However, topics were designated as optional so that regions with a need for data or interest in the topics would be able to select the specific topic module for inclusion in the CCHS (Cycle 2.1) in their own region. The advantage of this approach is that health regions can expand the health topic coverage tailored to the characteristics of the regions. The disadvantage is that, unlike the topic modules contained in the common content, the resulting data from the optional content modules is not easily generalized across Canada. Therefore, the size and characteristics of the regions in which the modules are used limit comparison of the results between regions.

Table 4.2 CCHS Cycle 2.1 Optional Topic Modules

• Alcohol dependence • Blood pressure check • Breast examinations • Breast self examinations • Colorectal cancer screening • Contacts with mental health professionals • Dental visits* • Depression • Dietary supplement use • Distress • Driving and safety* • Eating troubles assessment • Food choices • Health care system satisfaction • Health Status SF-36 • Health utility index* • Home safety • Illicit drug use • Leisure activities • Mastery	• Medication use* • Nicotine dependence • Oral health(2) – optional* • Patient satisfaction • Physical check-up • Physician counselling – smoking • Problem gambling • Prostate cancer screening • Satisfaction with availability of health care services • Satisfaction with life • Sedentary activities • Self-esteem • Smoking cessation aids • Social support • Stages of change (smoking) • Suicidal thoughts and attempts • Tobacco alternatives • Use of protective equipment • Work stress

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

5. Sample Design

5.1 Target Population

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

5.2 Health Regions

For administrative purposes, each province is divided into health regions (HR) and each territory was designated as a single HR (Table 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. Cycle 2.1 of the CCHS collected data in 123 HRs in the ten provinces, in addition to one HR per territory, totalling 126 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	17	24,280
Ontario	37	42,260
Manitoba	10	7,500
Saskatchewan	11	7,720
Alberta	9	14,200
British Columbia	16	16,090
Yukon	1	850
Northwest Territories	1	900
Nunavut	1	700
Canada	**126**	**130,700**

5.3 Sample Size and Allocation

To provide reliable estimates for these 126 HRs, and given the budget allocated to the CCHS cycle 2.1 component, a sample of 130,700 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 each HR sufficient sample with minimal disturbance to the provincial allocation of sample sizes. The sample sizes were enlarged before data collection to take into account out-of-scope and vacant dwellings and anticipated non-response. 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. The Yukon was attributed 850 sample units, 900 for Northwest Territory and 700 for Nunavut.

5.4 Frames, Household Sampling Strategies

The CCHS used three sampling frames to select the sample of households: 48% of the sample of households came from an area frame, 50% came from a list frame of telephone numbers and the remaining 2% 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[11]. 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 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) in which the average household income is high. 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 the number of households.

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

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. Where there is low population density, a three-step plan is used whereby two or three primary sampling units (PSU), which normally correspond to groups of EAs, are selected and dividing each PSU into clusters, six of which are sampled. The selection is made at each step using the PPS method.

Once the new clusters are listed, the sample is obtained using a systematic sampling of dwellings. 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 especially happens in sectors where there is an increase in the number of dwellings due to new construction, for example. 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 an operation, usually conducted at an aggregate level, is called sample stabilization. Moreover, one increases the required size of the sample by households to account for vacant dwellings, experience having shown that 13% of all dwellings are not occupied by households that are part of the field 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)	Yields (sampled households)
Toronto, Montréal, Vancouver	Cluster	200-250	6
Other cities	Cluster	150-200	8
Apartment frame	Apartment	Varies	5
Most rural areas / small urban centres	Enumeration area	300	10

Requirements specific to CCHS led to some modifications to this sampling strategy[12]. To get a base sample of 62,000 households for CCHS, 84,000 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 68,000 dwellings distributed across the various economic regions in Canada whereas the CCHS required a total of 84,000 dwellings distributed in the HRs, which have different geographic boundaries from those of the LFS economic regions. Overall, the CCHS required 24% more dwellings than those generated by the LFS selection

[12] 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.

mechanism, or an *adjustment factor* of 1.24 (84,000/68,000). At the HR level, however, the adjustment factors varied from 0.6 to 6.0, which required certain adjustments.

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 stabilisation, 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 of 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 of dwellings was repeated only once. Where the chosen approach created an unnecessary surplus of dwellings, stabilisation 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 new selected PSU requires a listing), and to avoid possible cluster effects created by too great a number of dwellings selected in a single PSU. This limit to the 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, percent Inuit and/or Aboriginal and median household income). There were five design strata in Yukon, ten in the Northwest Territory 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 as the one described above.

It is worth mentioning that the frame for CCHS covered 90% of the private households in Yukon, 97% in the Northwest Territory 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 8 HRs (the five 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 rates observed under the list frame approach were much higher than those for the RDD frame as they varied from 70% to 80%.

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 was only used in HR's where the area frame was the main source for the sample, the impact of the undercoverage of the list frame was minimal and was dealt with in weighting.

5.4.3 Sampling of Households from the RDD Frame of Telephone Numbers

In five 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[13]. 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 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 CCHS the hit rates ranged from 15% to 25% among the five HRs which required the use of the RDD frame.

5.5 Sampling of Interviewees

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[14]. The rule for selecting persons from households was defined as a function of the household composition by assigning varying probabilities of selection to individuals. Table 5.3 provides the probabilities of selection of those in the 12-19 age group versus the probabilities of selection of the 20-year-olds or over. For example, let us take a household of 3 persons: two adults over 20 and a 15-year-old teenager. The teenager would then have 5.8 more

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

[14] Béland, Y., Bailie, L., Catlin, G. and Singh, M.P. CCHS and NPHS – An Improved Health Survey Program at Statistics Canada, *2000 Proceedings of the American Statistical Association Meeting, Survey Research Methods Section*, Indianapolis: American Statistical Association, 677-682.

chances of being selected than the adults. Note that for the households which fall in "=" cells, individuals have equal probabilities of selection.

Table 5.3. Selection Strategy based on Household Composition – Probability of selection (12-19) versus probability of selection (20 or over)

Number of persons aged between 12 and 19	Number of persons aged 20 or over					
	0	1	2	3	4	5+
0	-	=	=	=	=	=
1	=	5.8x	4.8x	3.8x	2.8x	=
2	=	2.9x	2.4x	=	=	=
3+	=	=	=	=	=	=

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 12^{th} month is 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 2003, Q2: April and May 2003, Q3: June to August 2003 and Q4: September to November 2003). Within each collection quarter the selected dwellings were then 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 population in scope.

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 file of data.

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 Québec, 7 for Laurentides and 8 for Outaouais). 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 (Québec added 1,560 sample units, Laurentides added 1,630 and Outaouais added 1,910). Table 5.4 gives the sample allocation by sub-regional area.

Table 5.4. Final Sample Allocation including Extra Units in Quebec, Laurentides and Outaouais health regions

Sub-regional area	Total sample size (targeted)
Région de Québec	**2,458**
Charlevoix	307
De la Jacques-Cartier	308
Haute-Ville-Des Rivières	307
La Source	307
Orléans	307
Portneuf	308
Québec-Basse-Ville-Limoilou-Vanier	307
Ste-Foy-Sillery-Laurentien	307
Région de l'Outaouais	**2,528**
Hull	316
Grande-Rivière	316
Gatineau	316
Pontiac	316
Domaine des Forestiers	316
Vallée-de-la-Lièvre	316
Petite-Nation	316
Les Collines-de-l'Outaouais	316
Région des Laurentides	**2,363**
Jean-Olivier Chénier	324
Thérèse-de-Blainville	324
Hautes-Laurentides	339
Arthur-Buies	325
Pays-d'en-Haut	383
Trois-Vallées	343
Argenteuil	325

5.8 Special study on the mode of collection (CAPI versus CATI)

In order to better understand the differences caused by the methods of collection (CAPI and CATI) in the Canadian Community Health Survey, a special collection mode study was fully implemented as part of the CCHS cycle 2.1. Details of the mode study design are provided in Section 8 – Weighting as it had an impact on the CCHS cycle 2.1 weighting strategy.

The mode study used a split-plot design with a unique sample frame where the secondary sampling units were randomly assigned to either CAPI or CATI. The study was conducted between July and November 2003 in 11 sites selected to provide a good representation of each region in Canada. Each mode sample was allocated to the study sites proportionally to the CCHS cycle 2.1 sample sizes. Table 5.5 provides a detailed distribution of the mode study sample by selected site.

Table 5.5 – Mode Study Sample Sizes

Health Region	**CAPI**	**CATI**
St.John’s, Newfoundland and Labrador	135	100
Cape Breton, Nova Scotia	125	100
Halifax, Nova Scotia	200	150
Chaudière-Appalaches, Quebec	230	215
Montérégie, Quebec	405	390
Niagara, Ontario	235	230
Waterloo, Ontario	235	230
Winnipeg, Manitoba	320	320
Calgary, Alberta	350	290
Edmonton, Alberta	335	290
South Fraser, British Columbia	240	240
Total	2,810	2,555

The mode study sample was selected from one sampling frame only. In order to keep to a minimum the changes to the regular CCHS data collection procedures the sample was selected from the list frame of telephone numbers. The mode study made use of a stratified two-stage design where the 11 sites represented the study design strata. The first-stage units were the Census Sub-Divisions (CSD) while the telephone numbers were the second-stage units. Within each site, the sample of telephone numbers was selected as follows:

i. First stage: PPS-selection of CSDs;
ii. Allocation of the total sample (CAPI + CATI) of a given site to the sampled CSDs proportionally to their sizes;
iii. Second stage: Random selection of telephone numbers in each CSD.

Once the sample of telephone numbers was selected those cases for which a valid address was not available were tossed out of the process and added in the regular CCHS CATI sample; those cases were not part of the mode study but were fully integrated in CCHS telephone sample. Finally, controlling for the CSD within each study site, the telephone numbers with a valid address were assigned a method of collection (CAPI or CATI) on a random basis to constitute the two mode samples. The field interviewers then received the mode study cases (between 20 and 60) early July in a separate assignment than their CCHS assignment to clearly identify them as they were instructed to conduct only personal interviews (CAPI). (Note that the regular CCHS procedures allow field interviewers for telephone interviews in special circumstances.) The CATI mode sample cases were simply added in the CCHS monthly CATI samples (July, August and September). (The CATI mode study sample was completely transparent to the call centers' interviewers.) The reader should note that the mode study cases were also part of the CCHS cycle 2.1 master file as well.

The paper[15] describing the results of the mode study as implemented as part of the CCHS cycle 2.1 will be published soon.

[15] St-Pierre, M. and Béland, Y. – Mode Effects in the Canadian Community Health Survey: a comparison of CAPI and CATI, *2004 Proceedings of the American Statistical Association Meeting, Survey Research Methods Section,* Toronto: American Statistical Association, (to be published).

6. Data Collection

6.1 Questionnaire Design and Data Collection Method

The CCHS Cycle 2.1 questionnaire was administered using computer-assisted interviewing (CAI). Sample units selected from the telephone list frame were interviewed from centralised call centres using CATI. Units selected from the area frame were interviewed by decentralised field interviewers. In some situations, field interviewers were permitted to complete some or part of an interview by telephone.

CAI offers a number of data quality advantages over other collection methods. First, 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.

Second, 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.

Third, questions that are not applicable to the respondent are skipped automatically.

6.2 Supervision and Control

CATI interviewers worked in call centres in centralised offices and were supervised by a senior interviewer located in the same office. Transmission of cases from each of 5 CATI offices to head office was the responsibility of the regional office project supervisor, senior interviewer and the technical support team.

CAPI interviewers worked independently from their homes using laptop computers and were supervised from a distance by senior interviewers. Completed interviews were transmitted daily to Statistics Canada's head office using a secure telephone transmission directly from the interviewer's home.

An automated call scheduler, ie. a central system to optimise the timing of call-backs and the scheduling of appointments, was not available to support CATI collection. Instead, at the start of each month a batch of cases was assigned to each personal computer in each CATI office. The caseload on each PC was then managed manually. This approach was reasonably efficient and the absence of a call scheduler is not thought to have had an adverse effect on data quality.

6.3 Field Tests

Separate CAPI and CATI field tests were conducted during the late summer of 2002. The tests were conducted in Alberta and Québec.

The main objectives of the CAPI test were to evaluate respondent reaction to the questions and to obtain estimates of completion times for the various sections of the questionnaire. Field operations procedures, interviewer training and the computer-assisted interviewing application were also tested.

The objectives of the CATI test were similar to those of the CAPI test. In addition, the technical infrastructure of the CATI offices and procedures unique to CATI interviewing were tested.

6.4 Interviewing

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. Flags on the microdata file indicate whether a case was selected from the area frame or from a telephone frame (SAMC_TYP) and whether the interview was completed face-to-face, by telephone or using a combination of the two techniques (ADMC_N09).

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. Therefore, every effort was taken to keep proxy interviews to a minimum. The variable ADMC_PRX indicates whether a case was completed by proxy.

6.5 Minimising Non-response

Prior to the first contact by an interviewer, an introductory letter and brochure were delivered to each selected dwelling. These explained the importance of the survey and provided examples of how CCHS Cycle 2.1 data would be used.

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 no one was home, numerous call-backs were made. 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. During the final months of data collection, non-response cases were again approached and encouraged to participate in the survey. This diligence in contact may have resulted in stronger survey results by maximising the response rate.

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, Inuktitut and Cree.

At the end of data collection, a national response rate of 80.7% was achieved. Complete details regarding the response rates can be found in Section 9.

6.6 Privacy

To ensure the quality of the data collected, every effort was made to conduct interviews in private. In some situations, the respondent allowed another person to be present. Flags on the microdata file indicate whether somebody other than the respondent was present during the interview (ADMC_N10) and whether the interviewer felt that the respondent's answers were influenced by the presence of the other person (ADMC_N11).

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 2.1) 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 provides details on how these more complex variables were derived.

7.4 Weighting

The principle behind estimation in a probability sample such as the CCHS Cycle 2.1 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 microdata file, 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 Suppression of Confidential Information

It should be noted that the public use microdata file described above differs in a number of important respects from the survey 'master' file held by Statistics Canada. These differences are the result of actions taken to protect the anonymity of individual survey respondents. Protection of respondents is assured through suppression of individual values, variable grouping, and variable capping. Users requiring access to information excluded from the microdata files have three options: purchase custom tabulations; the Research Data Centres Program[16], and the remote access service. (See sub-section 12.3)

[16] The most current information about the Research Data Centres can be found at http://www.statcan.ca/english/rdc/index.htm

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 2.1) 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 (I1). 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, 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 A1 – Sample increase A2 – Stabilization A3 – Removal of out-of-scope units A4 – Household nonresponse A5 – Creation of person level weight A6 – Person nonresponse Final area frame weight ⇘	T0 – Initial weight T1 – Number of months T2 – Removal of out-of-scope numbers T3 – Coverage of the list frames T4 – Combination of the list frames T5 – Household nonresponse T6 – Households without telephone 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 2.1) 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 population size (based on 1991 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 2.1) 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)[17].

A1 – Sample increase

Some modifications were made to the default LFS mechanism at the time of sample selection for CCHS (Cycle 2.1). The LFS design provides approximately 68,000 dwellings nationally, while CCHS (Cycle 2.1) 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. As this modification boosted the sample size, adjustments to the weighting were needed in order to reflect the actual probability of selection. An adjustment factor,

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

A1, representing the 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 2.1). 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 subsampling 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 2.1). These dwellings were simply removed from the sample, leaving only a sample that consisted of in-scope dwellings. They 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 nonresponding households were distributed using response propensity classes to the responding households. The CHAID (Chi-Square Automatic Interaction Detector) algorithm available in Knowledge Seeker[18], 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 nonrespondents 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 collection period was the most significant characteristic (with 4 periods: January to March, April to June, July to September, and October to December) for the creation of classes in each HR. The rural/urban indicator was also significant for a small number of HRs. An adjustment factor was calculated within each class as follows:

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

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

Weight A3, for responding households, was multiplied by this factor to produce weight A4. Nonresponding 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 is larger for persons belonging to the 12-19 year old age group, for those households consisting of persons belonging to both the 12-19 and 20+ age groups (see Section 5.5 for more details). For the other households, this probability is equal to the inverse of the number of persons aged 12 and over, no matter which person is selected.

A6 – Person non-response

A CCHS (Cycle 2.1) 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 overcome this nonresponse. As for the treatment of household nonresponse, the adjustment was applied within classes based on characteristics available for both respondents and nonrespondents. All characteristics collected when rostering all household members were in fact available for the creation of the classes. The CHAID algorithm 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: sex, age group, urban/rural indicator, number of persons in the household, education, marital status and collection period 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 of responding persons was therefore multiplied by the above adjustment factor to produce weight A6. Nonresponding persons are 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 2.1) weight.

8.1.2 Weighting of the telephone frame sample

As mentioned previously, the telephone frame is composed of two frames: an RDD frame and a list frame, of which only one can be used in a specific HR. The list frame is always used as a complement to the area frame while the RDD frame is always used alone. However, units coming from these two frames are treated together and therefore are subject to the same adjustments. There are three exceptions: first, since the probability of selection 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 from sample selection. This adjustment was incorporated in the 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, the selection of telephone numbers is done within each RDD stratum. A RDD stratum is an aggregation of area code prefixes (ACP: the first six digits of a 10-digit number), each containing valid banks of one hundred numbers (see Norris and Paton[19] 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, within the HR for which the unit is selected. Hence, the probability of selection corresponds to the ratio of the number of sampled units to the number of telephone numbers in the list within the HR. Since the telephone frame sampling was done on a monthly basis (see adjustment T1) and because the list frame was updated twice 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.

[19] Norris, D.A. and Paton, D.G. (1991). Canada's Genereal 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 monthly 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 the monthly sample represented among the total sample. Since the coverage is different from one version of the list frame to another, the adjustment was done separately for the three versions which means that the sample of each version represented the total population. Thus, at this point, the total list frame sample represents about three times the total population. To correct this situation, samples from the three 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 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 2.1) 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 of each version of the list frame represent the entire population of the HR where the list frame was used. They all 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:

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

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 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:

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

The weight T4 of responding households was multiplied by this factor to produce the weight T5. Nonresponding 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. Similarly to T3, the telephone frame sample weights were inflated based on proportions observed using the area frame data, adjusting the weights for the undercoverage 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 for 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 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. Nonresponding 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. 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 2.1) 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 created until now, into a single weight, by applying a method of integration[20]. An adjustment factor between 0 and 1 was determined in such a way that it represented the relative importance of each

[20] 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.

sample in the total sample. This relative importance was measured in terms of sample size and design effect. The larger the proportion a sample represented in the total sample was, the higher was its relative importance in the total sample. For the design effect, the relative importance was bigger 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:

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

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 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 (I2)

The CCHS (Cycle 2.1) had initially planned 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 using CCHS (Cycle 2.1) data. 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 2.1) 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.

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 that, 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 2.1) weight was the 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 for Alberta, the post-stratification was done using a revised geography that contained 9 regions instead of the 17 initially used at the design stage and throughout data collection. Note also that the post-stratification was done at the CLSC region level for the three Quebec regions (2403, 2407 and 2415) that paid to get extra sample.

The population estimates were based on the 1996 Census counts and counts of birth, death, immigration and emigration. The average of these 2003 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 2.1) weight*** that can be found on the data file with the variable name WTSC_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 ten 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. Each territory was initially stratified into groupings of communities, where one community was selected within each group. Note that the capital of each territory formed a stratum on its own, and was consequently automatically selected at this first stage. This particularity only 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, let us first mention 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 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 poststratified 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, the other communities having been covered only by the area frame. The second detail relates to the seasonal adjustment. Because a strong concentration of the interviews was conducted during a short period of time in the Nunavut territory, the seasonal effect adjustment could not be applied efficiently. Estimations produced for the Nunavut using these weights will therefore not account for a possible seasonal effect in the data.

8.1.7 Particular aspects of the weighting in the HRs that are part of the collection mode effect study

A special treatment is administered to the units that come from the 11 HRs where the collection mode study took place. For those HRs, sample comes from three sources: area frame, regular telephone frame and collection mode study. Every sample has to be integrated together in order to form only one sample. Note that the weight specifically computed for the collection mode study analysis could not be used for the regular sample because parameters of CCHS are very different than parameters of the collection mode study.

Since collection mode study units come from the list frame, they are processed with the telephone frame. However, the regular weighting process has to be adapted for the collection mode study. Indeed, the telephone frame units are treated apart from the collection mode units until the two samples are integrated. This integration is done just before the household nonresponse adjustment of the telephone frame. The differences in the weighting process of the telephone frame for the 11 HRs where the collection mode study took place are described in the following section.

T0 –Initial weight

Contrary to the regular telephone frame sample, telephone numbers that were selected for the collection mode study were not drawn according to a simple random design within a HR. A sample of Census subdivisions (CSD) was first selected and then among the selected CSDs, a sample of telephone numbers was drawn. This extra stage was added in order to make sure that in-person interviewers didn't have to cover the entire HR. This particular aspect related to the collection mode study units is taken into account in the computation of initial weight.

T1 – Number of months

The adjustment for the number of months is not applied to collection mode study units. Contrary to the telephone frame sample that is drawn every month, only one sample was drawn for the collection mode study. The adjustment is therefore useless for those units. Also, since the collection mode study sample replaces the sample that was originally planned for the telephone frame, the number of months for which the sample comes from the list frame goes from 11 to 8 for those HRs.

T2 – Removal of out-of-scope numbers/units

For the collection mode study, the notion of out-of-scope units differs from a collection mode to another. Indeed, interviewers for the telephone component simply compose the selected telephone number and if the unit is in-scope, they conduct the interview. On the other hand, in-person interviewers go to the address that was associated with the selected telephone number at the moment the list frame was created. They conduct the interview if the dwelling is in-scope for the survey no matter if the telephone number of the household living at the address is the same as the one that was selected or not. It is possible that the telephone number that was selected is now out of service but there is still an in-scope household living at the associated address. If all in-person interviews had been conducted by telephone, the number of out-of-scope units would have increased and the loss of weight imputable to the removal of out-of-scope units would have increased as well. This low loss of weight due to the out-of-scope removal of the in-person component results in a surplus of weight and hence an overestimation of the actual coverage of the telephone frame. An adjustment is applied to the units for which the interview was conducted in person in order to make sure that their loss of weight (in proportion) is the same as the one of the units of the telephone component within a HR. That way, we make sure that we measure the actual coverage of the telephone frame at the moment of its creation.

T3 – List frame coverage

No change has been made to this step. The adjustment is applied to collection mode study units as well as regular sample units.

T4 - Combination of list frames

This adjustment applies only to the units that come from the regular telephone frame sample. It is not applicable to the collection mode study units since only one version of the list frame was used for the study.

T4b – Integration of the collection mode effect study

Up to this point, the regular telephone frame sample and the collection mode study sample were treated separately. This step consists in integrating the weights of those two components into a single weight. In order to integrate both samples, the following adjustment is computed for each component (regular sample and study) within each HR:

$$\frac{\textit{Sample size of the component}}{\textit{Total sample size of both components}}$$

This factor, multiplied by the weight T4, gives the weight T4b. From this point, weighting of the HRs where the collection mode effect study took place is exactly the same as weighting of other HRs. The subsequent adjustment are therefore applied exactly as described in section 8.1.2.

9. Data Quality

9.1 Response Rates

In total and after removing the out-of-scope units, 166,222 households were selected to participate in the CCHS (Cycle 2.1). Out of these selected households a response was obtained for 144,836 which results in an overall household-level response rate of 87.1%. Among these responding households 144,836 individuals (one per household) were selected to participate in the CCHS (Cycle 2.1) out of which a response was obtained for 134,072 which results in an overall person-level response rate of 92.6%. At the Canada level, this would yield a combined response rate of 80.7% for the CCHS (Cycle 2.1). Table 9.1 provides combined response rates as well as relevant information for calculation of them by health region or combined health region.

The CCHS (cycle 2.1) design in Alberta was done in July 2002 using the 17 health region boundaries that existed at that time. On April 1st, 2003, the government of Alberta reorganized their health region boundaries. That reorganization resulted in 9 new health regions which are now reported in the public use microdata file. As a result, it is not appropriate to report response rates for the new regions.

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.

$$\text{HHRR} = \frac{68{,}966 + 75{,}870}{77{,}528 + 88{,}694} = \frac{144{,}836}{166{,}222} = 0.871$$

$$\text{PPRR} = \frac{64{,}656 + 69{,}416}{68{,}966 + 75{,}870} = \frac{134{,}072}{144{,}836} = 0.926$$

Combined response rate	=	0.871 x 0.926
	=	0.807
	=	**80.7%**.

Table 9.1 / Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases / Tous les cas
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	77528	68966	89.0	68966	64656	93.8	83.4	88694	75870	85.5	75870	69416	91.5	78.3	80.7
NL	Total	2357	2204	93.5	2204	2066	93.7	87.7	2303	2140	92.9	2140	1988	92.9	86.3	87.0
	10901	484	433	89.5	433	401	92.6	82.9	496	462	93.1	462	427	92.4	86.1	84.5
	10902	508	487	95.9	487	457	93.8	90.0	448	415	92.6	415	387	93.3	86.4	88.3
	10903	487	466	95.7	466	431	92.5	88.5	378	352	93.1	352	328	93.2	86.8	87.7
	10904*	878	818	93.2	818	777	95.0	88.5	981	911	92.9	911	846	92.9	86.2	87.3
PE	Total	1120	1024	91.4	1024	970	94.7	86.6	1352	1189	87.9	1189	1092	91.8	80.8	83.4
	11901*	677	627	92.6	627	601	95.9	88.8	846	742	87.7	742	678	91.4	80.1	84.0
	11903	443	397	89.6	397	369	92.9	83.3	506	447	88.3	447	414	92.6	81.8	82.5
NS	Total	2799	2548	91.0	2548	2392	93.9	85.5	3092	2751	89.0	2751	2564	93.2	82.9	84.1
	12901	432	411	95.1	411	388	94.4	89.8	448	391	87.3	391	369	94.4	82.4	86.0
	12902	309	284	91.9	284	271	95.4	87.7	427	392	91.8	392	361	92.1	84.5	85.9
	12903	383	367	95.8	367	350	95.4	91.4	399	372	93.2	372	343	92.2	86.0	88.6
	12904	440	397	90.2	397	377	95.0	85.7	399	345	86.5	345	327	94.8	82.0	83.9
	12905	484	440	90.9	440	400	90.9	82.6	542	477	88.0	477	446	93.5	82.3	82.5
	12906	751	649	86.4	649	606	93.4	80.7	877	774	88.3	774	718	92.8	81.9	81.3
NB	Total	2909	2675	92.0	2675	2517	94.1	86.5	2801	2548	91.0	2548	2412	94.7	86.1	86.3
	13901	514	469	91.2	469	439	93.6	85.4	480	441	91.9	441	412	93.4	85.8	85.6
	13902	522	488	93.5	488	463	94.9	88.7	479	425	88.7	425	397	93.4	82.9	85.9
	13903	572	515	90.0	515	490	95.1	85.7	445	416	93.5	416	390	93.8	87.6	86.5
	13904*	609	561	92.1	561	527	93.9	86.5	697	636	91.2	636	610	95.9	87.5	87.1
	13906*	692	642	92.8	642	598	93.1	86.4	700	630	90.0	630	603	95.7	86.1	86.3
QC	Total	14381	12663	88.1	12663	11874	93.8	82.6	21023	17118	81.4	17118	15725	91.9	74.8	78.0
	24901	776	717	92.4	717	698	97.4	89.9	612	536	87.6	536	500	93.3	81.7	86.3
	24902	840	778	92.6	778	739	95.0	88.0	756	676	89.4	676	635	93.9	84.0	86.1
	24903	1078	927	86.0	927	880	94.9	81.6	3621	2849	78.7	2849	2598	91.2	71.7	74.0
	24904	933	827	88.6	827	795	96.1	85.2	979	821	83.9	821	753	91.7	76.9	81.0
	24905	746	646	86.6	646	583	90.2	78.2	701	608	86.7	608	565	92.9	80.6	79.3
	24906	2117	1751	82.7	1751	1637	93.5	77.3	1727	1257	72.8	1257	1131	90.0	65.5	72.0
	24907	780	708	90.8	708	656	92.7	84.1	3474	2860	82.3	2860	2606	91.1	75.0	76.7
	24908	669	604	90.3	604	561	92.9	83.9	622	542	87.1	542	507	93.5	81.5	82.7
	24909	714	625	87.5	625	580	92.8	81.2	645	553	85.7	553	510	92.2	79.1	80.2
	24911	707	655	92.6	655	628	95.9	88.8	532	446	83.8	446	389	87.2	73.1	82.1
	24912	828	745	90.0	745	717	96.2	86.6	837	704	84.1	704	655	93.0	78.3	82.4
	24913	940	803	85.4	803	681	84.8	72.4	791	626	79.1	626	576	92.0	72.8	72.6
	24914	882	780	88.4	780	721	92.4	81.7	813	673	82.8	673	629	93.5	77.4	79.6

Table 9.1 Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases / Tous les cas
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
	24915	888	777	87.5	777	734	94.5	82.7	3327	2652	79.7	2652	2448	92.3	73.6	75.5
	24916	1483	1320	89.0	1320	1264	95.8	85.2	1586	1315	82.9	1315	1223	93.0	77.1	81.0
ON	Total	24760	21749	87.8	21749	20326	93.5	82.1	29701	25066	84.4	25066	22451	89.6	75.6	78.5
	35926	520	424	81.5	424	391	92.2	75.2	613	519	84.7	519	471	90.8	76.8	76.1
	35927	553	481	87.0	481	442	91.9	79.9	532	449	84.4	449	412	91.8	77.4	78.7
	35930	1057	912	86.3	912	841	92.2	79.6	1146	984	85.9	984	874	88.8	76.3	77.8
	35931	378	314	83.1	314	295	93.9	78.0	530	458	86.4	458	409	89.3	77.2	77.5
	35933	498	444	89.2	444	426	95.9	85.5	622	529	85.0	529	491	92.8	78.9	81.9
	35934	467	422	90.4	422	401	95.0	85.9	493	417	84.6	417	373	89.4	75.7	80.6
	35935	511	420	82.2	420	398	94.8	77.9	588	483	82.1	483	432	89.4	73.5	75.5
	35936	843	770	91.3	770	724	94.0	85.9	902	761	84.4	761	684	89.9	75.8	80.7
	35937	1111	894	80.5	894	811	90.7	73.0	1158	976	84.3	976	852	87.3	73.6	73.3
	35938	302	287	95.0	287	278	96.9	92.1	871	725	83.2	725	657	90.6	75.4	79.7
	35939*	391	357	91.3	357	332	93.0	84.9	1199	1030	85.9	1030	945	91.7	78.8	80.3
	35940	424	408	96.2	408	397	97.3	93.6	479	404	84.3	404	378	93.6	78.9	85.8
	35941	586	510	87.0	510	492	96.5	84.0	651	560	86.0	560	504	90.0	77.4	80.5
	35942	501	465	92.8	465	440	94.6	87.8	554	472	85.2	472	433	91.7	78.2	82.7
	35943	586	502	85.7	502	468	93.2	79.9	611	503	82.3	503	450	89.5	73.6	76.7
	35944	984	856	87.0	856	816	95.3	82.9	994	855	86.0	855	781	91.3	78.6	80.7
	35945	299	261	87.3	261	244	93.5	81.6	697	582	83.5	582	533	91.6	76.5	78.0
	35946	941	823	87.5	823	771	93.7	81.9	1101	926	84.1	926	831	89.7	75.5	78.5
	35947*	636	590	92.8	590	557	94.4	87.6	934	818	87.6	818	758	92.7	81.2	83.8
	35949	266	234	88.0	234	207	88.5	77.8	598	508	84.9	508	456	89.8	76.3	76.7
	35951	1212	1089	89.9	1089	1017	93.4	83.9	1334	1131	84.8	1131	1030	91.1	77.2	80.4
	35952	470	425	90.4	425	387	91.1	82.3	485	432	89.1	432	386	89.4	79.6	80.9
	35953	1521	1325	87.1	1325	1233	93.1	81.1	1503	1244	82.8	1244	1057	85.0	70.3	75.7
	35955	200	171	85.5	171	166	97.1	83.0	889	761	85.6	761	688	90.4	77.4	78.4
	35956	490	448	91.4	448	431	96.2	88.0	494	433	87.7	433	397	91.7	80.4	84.1
	35957	476	425	89.3	425	394	92.7	82.8	477	403	84.5	403	367	91.1	76.9	79.9
	35958	572	511	89.3	511	472	92.4	82.5	687	592	86.2	592	536	90.5	78.0	80.1
	35960	761	705	92.6	705	664	94.2	87.3	1012	866	85.6	866	784	90.5	77.5	81.7
	35961	640	591	92.3	591	543	91.9	84.8	676	589	87.1	589	519	88.1	76.8	80.7
	35962	540	497	92.0	497	479	96.4	88.7	635	548	86.3	548	488	89.1	76.9	82.3
	35965	934	812	86.9	812	757	93.2	81.0	1087	921	84.7	921	836	90.8	76.9	78.8
	35966	613	571	93.1	571	542	94.9	88.4	750	647	86.3	647	581	89.8	77.5	82.4
	35968	890	763	85.7	763	726	95.2	81.6	966	788	81.6	788	680	86.3	70.4	75.8

Table 9.1 Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases / Tous les cas
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
	35970	1096	980	89.4	980	928	94.7	84.7	1217	996	81.8	996	865	86.8	71.1	77.5
	35995	2491	2062	82.8	2062	1856	90.0	74.5	2216	1756	79.2	1756	1513	86.2	68.3	71.6
MB	Total	4137	3786	91.5	3786	3617	95.5	87.4	4810	4331	90.0	4331	4015	92.7	83.5	85.3
	46910	1242	1080	87.0	1080	1033	95.6	83.2	1478	1320	89.3	1320	1210	91.7	81.9	82.5
	46915*	732	681	93.0	681	650	95.4	88.8	865	789	91.2	789	744	94.3	86.0	87.3
	46920*	516	490	95.0	490	469	95.7	90.9	627	559	89.2	559	515	92.1	82.1	86.1
	46930	356	332	93.3	332	322	97.0	90.4	438	405	92.5	405	377	93.1	86.1	88.0
	46940	432	406	94.0	406	379	93.3	87.7	498	450	90.4	450	423	94.0	84.9	86.2
	46960*	859	797	92.8	797	764	95.9	88.9	904	808	89.4	808	746	92.3	82.5	85.6
SK	Total	4530	4104	90.6	4104	3887	94.7	85.8	4467	3965	88.8	3965	3700	93.3	82.8	84.3
	47901*	1140	1050	92.1	1050	973	92.7	85.4	945	846	89.5	846	793	93.7	83.9	84.7
	47904	740	652	88.1	652	620	95.1	83.8	678	598	88.2	598	554	92.6	81.7	82.8
	47905*	702	670	95.4	670	659	98.4	93.9	627	545	86.9	545	507	93.0	80.9	87.7
	47906	787	698	88.7	698	662	94.8	84.1	764	680	89.0	680	639	94.0	83.6	83.9
	47907*	720	662	91.9	662	622	94.0	86.4	582	520	89.3	520	489	94.0	84.0	85.3
	47909*	441	372	84.4	372	351	94.4	79.6	871	776	89.1	776	718	92.5	82.4	81.5
AB	Total	7930	7165	90.4	7165	6686	93.3	84.3	8846	7805	88.2	7805	7185	92.1	81.2	82.7
	48901	566	515	91.0	515	492	95.5	86.9	485	436	89.9	436	400	91.7	82.5	84.9
	48902	397	355	89.4	355	335	94.4	84.4	460	424	92.2	424	381	89.9	82.8	83.5
	48903*	675	612	90.7	612	576	94.1	85.3	708	628	88.7	628	584	93.0	82.5	83.9
	48904	1362	1209	88.8	1209	1135	93.9	83.3	1349	1149	85.2	1149	1063	92.5	78.8	81.1
	48906	577	531	92.0	531	487	91.7	84.4	543	469	86.4	469	436	93.0	80.3	82.4
	48907	460	437	95.0	437	419	95.9	91.1	464	409	88.1	409	382	93.4	82.3	86.7
	48908*	672	609	90.6	609	556	91.3	82.7	809	729	90.1	729	676	92.7	83.6	83.2
	48910	1252	1096	87.5	1096	1050	95.8	83.9	1250	1079	86.3	1079	990	91.8	79.2	81.5
	48911	414	378	91.3	378	351	92.9	84.8	453	397	87.6	397	371	93.5	81.9	83.3
	48912	570	516	90.5	516	495	95.9	86.8	401	361	90.0	361	335	92.8	83.5	85.5
	48913	440	403	91.6	403	319	79.2	72.5	366	332	90.7	332	299	90.1	81.7	76.7
	48914*	545	504	92.5	504	471	93.5	86.4	1558	1392	89.3	1392	1268	91.1	81.4	82.7
BC	Total	9713	8518	87.7	8518	7959	93.4	81.9	10056	8757	87.1	8757	8099	92.5	80.5	81.2
	59911	231	224	97.0	224	218	97.3	94.4	473	432	91.3	432	409	94.7	86.5	89.1
	59912	311	290	93.2	290	282	97.2	90.7	397	359	90.4	359	345	96.1	86.9	88.6
	59913	784	720	91.8	720	687	95.4	87.6	639	549	85.9	549	518	94.4	81.1	84.7
	59914	616	565	91.7	565	547	96.8	88.8	523	468	89.5	468	441	94.2	84.3	86.7
	59921	640	587	91.7	587	549	93.5	85.8	602	505	83.9	505	469	92.9	77.9	82.0
	59922	999	870	87.1	870	821	94.4	82.2	866	752	86.8	752	692	92.0	79.9	81.1

Table 9.1 / Tableau 9.1		Area frame / Base aréolaire							Phone frames / Bases téléphoniques							All cases / Tous les cas
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
	59923	1004	886	88.2	886	776	87.6	77.3	996	827	83.0	827	754	91.2	75.7	76.5
	59931	397	354	89.2	354	333	94.1	83.9	625	551	88.2	551	496	90.0	79.4	81.1
	59932	1061	854	80.5	854	808	94.6	76.2	921	788	85.6	788	711	90.2	77.2	76.6
	59933	689	604	87.7	604	547	90.6	79.4	600	522	87.0	522	475	91.0	79.2	79.3
	59941	852	742	87.1	742	692	93.3	81.2	702	585	83.3	585	538	92.0	76.6	79.2
	59942	562	472	84.0	472	433	91.7	77.0	539	477	88.5	477	437	91.6	81.1	79.0
	59943	320	310	96.9	310	302	97.4	94.4	628	580	92.4	580	553	95.3	88.1	90.2
	59951*	716	588	82.1	588	540	91.8	75.4	1045	935	89.5	935	859	91.9	82.2	79.4
	59952	531	452	85.1	452	424	93.8	79.8	500	427	85.4	427	402	94.1	80.4	80.1
Terr.	60901*	2892	2530	87.5	2530	2362	93.4	81.7	243	200	82.3	200	185	92.5	76.1	81.2

* = 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. as 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 2.1). 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 CCHS (Cycle 2.1) 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 of 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 one estimates that 25% of Canadians aged 12 and over are regular smokers and that this estimates is found to have 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 CCHS (Cycle 2.1) 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 microdata file. 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 microdata file 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 microdata files 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 microdata file.

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? ***(SMKC_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? (**SMKC_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 microdata file 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 **SMKC_204** and the weight **WTSC_M.** Next, sum this value over those records with a value of "daily" to the variable **SMKC_202**. The denominator ($\hat{Y}$) is obtained by summing the final weight of those records with a value of "daily" to the variable **SMKC_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 microdata file, 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 weighted 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 10 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	$0.0 \leq CV \leq 16.6$	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 microdata file 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 30, 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 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 CV's 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 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 CV's 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 ($\hat{X}$ -k, $\hat{X}$ +k), 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_X = {0.777 - (2) (0.777) (0.006) , 0.777 + (2) (0.777) (0.006)}

CI_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 = .004. Hence,

$$z = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}} = \frac{.194 - .162}{.004} = \frac{.032}{.004} = 8$$

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 public use microdata file. 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 public-use microdata file 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 **WTSC_M** represents the CCHS cycle 2.1 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 2.1 (SMKC_202):

Position 1-3:	**SMK**	**depression module**
Position 4:	**C**	**Cycle 2.1**
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:

ADM	Administration	LEI	Leisure activities
ALC	Alcohol	MAM	Mammography
ALD	Alcohol dependence / abuse	MAS	Mastery
BPC	Blood pressure	MED	Medication use
BRX	Breast examinations	MEX	Maternal experiences
BSX	Breast self examinations	NDE	Nicotine dependence
CCC	Chronic conditions	OH1	Oral health 1
CCS	Colorectal cancer screening	OH2	Oral health 2
CIH	Changes made to improve health	ORG	Voluntary organizations
CMH	Contacts with mental health professionals	PAC	Physical activities
CPG	Problem gambling	PAP	PAP smear test
DEN	Dental visits	PAS	Patient satisfaction
DHH	Demographics and household	PCU	Physical check-up
DIQ	Distress and mental health (Quebec)	PSA	PSA test
DIS	Distress	QMD	Medication use (Quebec)
DPS	Depression	RAC	Restriction of activities
DRV	Driving and safety	REP	Repetitive strain
DSU	Dietary supplement use	SAC	Sedentary activities
EDU	Education	SAM	Sample Identifiers
ETA	Eating troubles assessment	SCA	Smoking cessation aids
ETS	Exposure to second hand smoke	SCH	Smoking - stages of change
FDC	Food choices	SDC	Socio-demographics
FIN	Food insecurity	SFE	Self-esteem
FLU	Flu shots	SFR	Health status - SF-36
FVC	Fruit and vegetable consumption	SMK	Smoking
GEN	General health	SPC	Smoking - physician counselling
GEO	Geographic identifiers	SSM	Social support
HCS	Health care system satisfaction	SUI	Suicidal thoughts and attempts
HCU	Health care utilization	SWA	Satisfaction with availability
HMC	Home care	SWL	Satisfaction with life
HMS	Home safety	SXB	Sexual Behaviour
HUI	Health Utility Index (HUI)	TAL	Tobacco alternatives
HWT	Height and Weight	TWD	Two-week disability
IDG	Illicit drugs	UPE	Use of protective equipment
INC	Income	WST	Work stress
INJ	Injury	WTS	Sample weights
INS	Insurance coverage	YSM	Youth smoking
LBF	Labour force		

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

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, microdata files 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 microdata file 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 microdata files 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 2.1 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.