TABLE OF CONTENTS

Page

List of Appendices

1. Introduction

The National Population Health Survey (NPHS) collects information related to the health of the Canadian population and related socio-demographic information. The NPHS is composed of three components: the household survey, the Health Care Institution Survey and the Northern Territories survey. These Public Use Microdata Files (PUMF) contain data collected in the household component of NPHS Cycle 3, 1998-1999.

The NPHS household component includes household residents in all provinces, with the exclusion of populations on Indian Reserves, Canadian Forces Bases and some remote areas in Québec and Ontario. The first Cycle of data collection began in 1994 and data will be collected every second year, for approximately 20 years in total. Three cycles of collection are now completed for each component: NPHS Cycle 1 (1994-1995), NPHS Cycle 2 (1996-1997) and NPHS Cycle 3 (1998-1999).

For the first cycle, a sample of approximately 20,000 households was drawn from the Labour Force Survey sampling frame. For Cycle 3, this frame was also used to select an additional sample of recent immigrants and young children, thus ensuring that the data represent the 1998-1999 Canadian population.

NPHS collects general health information from all household members and, in each household, a person, randomly selected during cycle 1 answers a more in-depth interview on health questions. For Cycle 3, approximately 49,000 respondents answered the general portion of the questionnaire while approximately 17,000 answered the more detailed health portion.

The questionnaire includes questions related to health status, use of health services, determinants of health, chronic conditions and activity restrictions. The use of health services was measured through questions on visits to health care providers, both traditional and non-traditional, hospital cares and on use of drugs and other medications. Health determinants that are explored include smoking, alcohol use and physical activity. New content for the third Cycle of NPHS includes family medical history, self-care and nutrition. The socio-demographic information collected includes age, sex, education, ethnicity, household income and labour force status

This document has been produced to facilitate the manipulation of the two 1998-1999 cross-sectional microdata files containing the results from NPHS Cycle 3 (1998-1999). These files are described in more detail in Chapter 4 and in the appendices.

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

- For technical/general data support call:
  Electronic Products Help Line: 1-800-949-9491

- For custom tabulations/general data support call:
  Client Custom Services, Health Statistics Division: 1-613-951-1746
  E-mail: hd-ds@Statcan.ca

- For remote access support call:
  Colette Koeune 1-613-951-1653
  E-mail: nphs@statcan.ca
  Fax: 1-613-951-4198

- For survey content support call:
  Mario Bédard 1-613-951-8933
  France Bilocq 1-613-951-6956
  Fax: 1-613-951-4198

2. Background

In the fall of 1991, the National Health Information Council (NHIC) recommended that an ongoing national survey of population health be conducted. This recommendation was based on consideration of the economic and fiscal pressures on the health care systems and the commensurate requirement for information with which to improve the health status of the population in Canada. Existing sources of health data were unable to provide a complete picture of the health status of the population and the myriad factors that have an impact on health.

Commencing in April 1992, Statistics Canada received funding for development of a National Population Health Survey. The survey was designed to be flexible and to produce valid, reliable and timely data. Also, it was to be responsive to changing requirements, interests and policies.

3. Objectives

The objectives of the NPHS are to:

- aid in the development of public policy by providing measures of the level, trend and distribution of the health status of the population;
- 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 utilisation, including alternative as well as traditional services;
- provide information on a panel of people who will be followed over time to reflect the dynamic process of health and illness;
- provide the provinces and territories and other clients with a health survey capacity that will permit supplementation of content or sample;
- allow the possibility of linking survey data to routinely collected administrative data such as vital statistics, environmental measures, community variables, and health services utilisation.

4. Survey Content

The objectives described in Chapter 3 provided a broad direction for the NPHS, particularly concerning the type of information to be collected. The first section of this chapter discusses the general criteria used for the selection of survey content and gives a broad summary of sections and changes. The next section describes detailed changes to existing content for the third cycle of NPHS. The next section focuses on the variables from the previous cycles that were fed back and used in cycle 3. The last section details new content for the 1998-1999 survey.

4.1 Selection Criteria

Survey content was selected according to the following criteria:

1) Information should relate to, and help monitor, the health goals and objectives of the provinces and territories. Where health goals have not been established, for example, at the national level, policies and programs could be considered in the selection of survey content;

2) The information should not duplicate data available from other sources;

3) With a view to increasing the understanding of health and its determinants, information collected should provide new knowledge in areas that have not been adequately studied;

4) The survey should focus on behaviours or conditions amenable to prevention, treatment or intervention;

5) The survey should collect information about conditions that impose the greatest burden, in terms of suffering or cost, on affected individuals, the general population or the health care systems;

6) The survey should collect information on factors related to good health, not just those related to illness.

In each household, some limited information was collected from all household members (general component) and one person in each household was randomly selected for a more in-depth interview (health component). Reflecting the above criteria, the questionnaire included sections on health status, use of health services, risk factors and demographic and socio-economic status. For example, health status was measured through questions on self-perception of health, functional ability, chronic conditions and activity restriction. The use of health services was measured through questions on visits to health care providers, hospital care and drug use. Behavioural risk factors included smoking, alcohol use and physical activity. In addition, focus content for this third cycle of the survey covered the following topics: self-care, family medical history and nutrition. Senses of coherence questions from cycle 1 were also

re-administered. Demographic and socio-economic information included age, sex, education, ethnicity, household income and labour force status.

In order to reduce the collection costs, the 1998-1999 Health Promotion Survey (HPS) questions were integrated into the NPHS rather than being done as a separate supplemental survey as in 1994-1995. Additional questions covered tobacco alternatives.

The complete 1998-1999 NPHS questionnaire is provided in Appendix A.

4.2 1998-1999 Changes to Existing Content

The major enhancement from 1996-1997 was to incorporate the content for children under 12 years of age into the regular flow of the questionnaire.

General Component - changes by section:

Restriction of Activities:

RESTR-Q1D	Added «or school».

Socio-demographic Characteristics

SOCIO-Q6	Added an interviewer instruction about babies.

Education:

EDUC-Q7	Added the word "ever" to ensure that the cumulative highest level of education is understood.

Income:

INC-Q1	The skip for "none" to Q1 was changed to go to Q3 instead of out of the section.
INC-Q3	Exact household income was asked first. The cascade approach was used only in cases of non-response.

Health Component - changes by section:

The 1996 Alberta supplement content was dropped. The focus content on Access to Services was dropped with the exception of the core questions on mammography, PAP smear and blood pressure.

Preventive Health:

Blood pressure, mammography and PAP smear test questions were put back together under Preventive Health - as in 1994. They had been separated out in 1996 so that they would fit with the focus questions on access to services. All women’s heath questions were placed together in this section including questions on pregnancy and birth of last child.

Insurance:
All questions about insurance were placed together in this section.

Injuries:

IN-Q7	In 1994 this question was only asked of those currently working. This was an incorrect assumption, in fact people may not currently be working because of the injury. In 1998 it was asked of everyone 15 years and older regardless of working status.

Drugs:

DRG-Q3	Added a follow-up question for each drug taken as to whether or not it was a prescription drug.

Social Support:
Replaced the whole section with a new measure of social support (Medical Outcomes Study (MOS). See documentation on derived variables.

Agreement to Share:
Changed the wording of the agreement to share question in order to accommodate all survey cycles. Changed the words «from this interview» to «from all interviews».

4.3 Data Feedback and Follow-up Questions

In order to reduce respondent burden, questions, which had been asked in previous cycles and to which the answers would not change over time (e.g., country of birth) were not asked in NPHS Cycle 3. For variables that could change over time but only if certain actions had occurred (e.g., level of education), updating was only done if appropriate. Some answers from earlier cycles were fed back into the cycle 3 interview. This proved to be a valuable tool resulting in better quality estimates.

Restriction of Activities:
Whether or not the respondent had a disability in 1996 was used. If the status changed, an explanation of that change was probed.

Chronic Conditions:
For all respondents, selected chronic conditions (asthma, arthritis, high blood pressure, migraine headaches, diabetes, epilepsy, stomach or intestinal ulcers and the effects of a stroke) were fed back in an attempt to help explain change. If it was a newly acquired condition then the date of onset of the condition was captured.

Socio-demographic Characteristics:
For all respondents, flags indicating that country of birth and ethnic origin had been collected were re-input. Since the response categories to race were changed, this variable was re-collected. Language first learned and still spoken was asked again because this can change over time.

Education:

For all respondents, a flag indicating the highest level of education was re-input. Screening questions determined if the respondent was currently attending a learning institution between cycles. If so, educational attainment was collected anew.

Labour Force:

For all respondents, the employer name, type of industry and duties for the main job in 1996-1997 were fed back. If the respondent indicated that they worked in the previous year, they were asked to confirm the employer name.

Blood Pressure, Mammography, Pap Smear Test and Smoking:

In cycle 1 and cycle 2 the respondent was asked if he or she «ever» had his or her blood pressure taken (or ever had a mammography etc.). There were conflicting data between the two years. In 1998, the questions were repeated; however, when the respondent said that he or she has not had it taken and in the previous cycle reported that it had been taken (or had one taken etc.), then a probing question was asked.

Health Number:

A flag indicating if the health number collected in a previous cycle was «valid» was used. If the respondent's health number had not changed since last cycle and was invalid then the health number was asked again.

4.4 New Content for NPHS Cycle 3 (1998-1999)

General Component:

Health Care Utilisation:

Added a write-in question to get reason for seeking health care in the United States.

Change of Residence

For top-up purposes, this section was added to identify place of residence for new members of longitudinal households since October 1, 1994.

Personal Income

In the general, exact personal income was asked of the respondents aged 15 years and over. This was followed by the cascade approach in cases of non-response to the exact income question.

Food Insecurity

Three screening questions were used to identify those households «at risk» for food insecurity.

Health Component:

Hysterectomy:

Questions on prevalence age of hysterectomy and reason for having it were added.

Self-care:

This is focus content in 1998-1999.

Family Medical History:

This is focus content in 1998-1999.

Nutrition:

These questions were added in order to have some baseline nutrition questions. Nutrition was suggested as a focus content in 1998-1999. A paper and pencil pilot test was conducted in the fall of 1997. Nutrition questions will continue to be developed and possibly be included as focus content in a future cycle.

Sense of Coherence:

Questions on sense of coherence were first asked in 1994 (Cycle 1). These questions will appear on the questionnaire periodically because a person's sense of coherence is said to be relatively stable after the age of 35. The data collected will test that hypothesis. In cycle 3 data on sense of coherence were collected for the first time for the top-up sample and for persons who were less than 18 in 1994.

Smoking:

SMK-Q5A	Volume asked of former smokers
SMK-Q5B	Volume and frequency asked of occasional smokers

Tobacco Alternatives:

These Health Promotion questions are to be used as baseline questions on cigar, pipe, chewing tobacco and cigars.

Some Health Promotion questions were retained as part of core questions and will now be asked each cycle.

5. Sample Design

The target population of the NPHS includes household residents in all provinces, with the principal exclusion of populations on Indian Reserves, Canadian Forces Bases and some remote areas in Québec and Ontario.

5.1 1998–1999 Sample Design

5.1.1 Sample Design for the Core Household Component

In the first cycle of the NPHS, the sample was created by first selecting households and then within each household choosing one member to be the longitudinal respondent. For the third cycle, the distinction is made between the samples selected for longitudinal and for cross-sectional purposes.

The longitudinal sample for 1998–1999 consists of all longitudinal respondents chosen in cycle 1 who had completed at least the general component of the questionnaire in 1994–1995. This included 2,022 persons who were under the age of 12 in cycle 1 (previously interviewed as part of the 1994–1995 National Longitudinal Survey of Children and Youth (NLSCY) who were included in the NPHS sample for 1996–1997). Units selected in 1994–1995 as part of supplemental buy-in samples were excluded. Only the longitudinal respondent was traced using contact information collected in previous cycles; no attempt was made to follow all household members over time. For cross-sectional purposes, all household members currently living with the longitudinal respondent were interviewed. The selected longitudinal respondent's data will be used for longitudinal purposes and cross-sectional purposes.

The core sample selected in 1994–1995 was increased for 1998–1999 cross-sectional estimates (see 5.1.2). This supplemental sample was taken to compensate for sample attrition and to account for initially absent persons (i.e., infants and new immigrants). Overall, the cross-sectional sample in cycle 3 was slightly larger than the cycle 1 sample.

Longitudinal Core Sample Size by Current Province
Number of Persons

Province	1994–1995 Core Effective Sample [1]	1998–1999 Core Sample	% of 1994–1995 Core Sample Followed
Newfoundland	1,212	1,082	89.3
Prince Edward Island	1,184	1,037	87.6
Nova Scotia	1,271	1,085	85.4
New Brunswick	1,277	1,125	88.1
Québec	3,430	3,000	87.5
Ontario	5,335	4,307	80.7
Manitoba	1,346	1,205	89.5
Saskatchewan	1,320	1,168	88.5
Alberta	1,697	1,544	91.0
British Columbia	2,023	1,723	85.2
Total	20,095	17,276	86.0

Since no new longitudinal units were selected in cycles 2 and 3, the population covered by the longitudinal sample in 1998–1999 is 4 years old and older. That is to say, no selected person is aged from 0 to 3 in cycle 3, and therefore, no direct estimates of variables contained on the core health component of the questionnaire will be possible for these ages. Another implication of not selecting any new longitudinal units in cycle 3 is that people who immigrated to Canada in the last four years (since the last time longitudinal units were selected in cycle 1) are not represented by the selected persons forming the longitudinal sample. However, because a supplemental sample was drawn on these sub-populations, estimates can be made on the entire 1998–1999 cross-sectional population. Any 0 to 3 year olds in the households of longitudinal respondents, as well as recent immigrants who joined these households, are also represented by the core sample.

The target population in 1994–1995 was household residents in all provinces excluding persons living on Indian Reserves, Canadian Forces Bases and remote areas in Ontario and Québec. Cross-sectionally, the core target population includes all household members currently living with longitudinal respondents. The target population for the supplemental sample is described in section 5.1.2.

To identify whether persons were part of the target population or not, the following rules were developed for data collection purposes.

1 Effective sample, excluding dwellings that were either non-eligible, vacant, under construction or rejected by the rejective method.

Status of Longitudinal Respondent	**Action Taken**
Dead	For longitudinal respondents identified to be deceased at the time of the cycle 3 interview, the death was confirmed against the Canadian Mortality Database. Longitudinal respondents who have died are part of the longitudinal file, but do not appear on cross-sectional files.
Moved into an institution	Longitudinal respondents who moved from a private household to a health care institution were interviewed by the Institutional component of the NPHS. Their data appear on the longitudinal household file but they are not part of the cross-sectional file.
Moved to Northwest Territories or Yukon	For longitudinal respondents who moved to the North, attempts were made to collect their new location. When possible, the longitudinal respondent was interviewed using the household questionnaire. They are part of the longitudinal file, but do not appear on the cross-sectional files.
Moved to Indian Reserve Or Canadian Forces Base	If a longitudinal respondent was traced to an Indian Reserve or Canadian Forces Base, attempts were made to interview this selected person and the other members of his household using the household questionnaire. They are part of the longitudinal file but do not appear on the cross-sectional file.
Moved out of Canada Temporarily	In cases where the longitudinal respondent moved out of Canada for a period of time but is expected to return to living in Canada, attempts were made to interview the longitudinal respondent. Again, these units form part of the longitudinal file, but do not appear on the cross-sectional file.
Moved out of Canada Permanently	If the longitudinal respondent moved out of Canada and is not expected to return, her new location was collected and may be used to follow up in future waves. These persons were not interviewed and do not appear on the longitudinal or cross-sectional files.

5.1.2 1998–1999 Sample Design for the Cross-sectional Sample Supplement

Two small supplemental or «top-up» samples were drawn to make the overall cross-sectional sample more representative of the 1998–1999 population. The

first was designed to take into account the «initially absent population», those persons not available to be sampled in 1994–1995. The second was taken to compensate for sample attrition since cycle 1. As in the core survey, a general component of the questionnaire was administered to all of the members of a responding household. As well, one member of the household was selected to answer the health component. These supplemental samples are combined with the core sample to produce one large cross-sectional file of data.

The first supplemental sample was a random sample of the initially absent population, infants born in 1995 and thereafter and immigrants who entered Canada since the beginning of 1995. Four different rotation groups from the Labour Force Survey (LFS) were used as the frame for these populations, one for each the NPHS collection periods. When each rotation group reached its last month in the LFS sample, a supplement was added to the LFS questionnaire asking country of birth of each member of the household. Using the response to this question, as well as the household roster, all new immigrants and infants were found, using January 1, 1995, as the cut-off date.

All 422 immigrant households were surveyed by the NPHS; three or four months after the corresponding rotation group had rotated out of the LFS. Infants were sampled from only two of the rotation groups and interviewed only in quarters 2 and 3, due to a scheduling conflict with the NLSCY. A total of 758 infant households were surveyed, roughly half the total available from the two rotation groups. In quarter 3, the NLSCY needed a small sample of 0-year-olds. The NLSCY and not the NPHS thus interviewed a portion of these infants. A special weight stabilisation procedure was carried out to account for this special case, as well as for households with several children.

Since the household roster was known in advance from the LFS interview, the person to be administered the health questionnaire was chosen before the cases were sent to the field and tracing was carried out if this selected person had moved in the three or four months since the LFS interview. All interviews of immigrant households were initially scheduled to be personal visits, to assist in the tracing process, while all interviews of the infant households were done by telephone, to reduce costs.

The second supplemental sample was taken to replace the sample that had been lost to attrition since cycle 1. Individuals in dwellings that were part of the original sampling frame but whose household members did not respond in 1994–1995 were eligible. A total of 2,598 households were contacted and asked to participate. Since only street address information was available for the associated dwelling, all interviews were scheduled for personal visits. Interviewers filled in the household roster at the first contact. Due to a problem in the data collection computer application (i.e., the questionnaire on computer), no 0-to-11-year-olds were selected to be administered the health questionnaire in this supplemental sample, which was adjusted for in the weighting procedures.

1998–1999 Cross-sectional Sample Size by Current Province
Number of Households

Province	Supplemental Sample	Core Sample	Total
Newfoundland	140	1,082	1,222
Prince Edward Island	167	1,037	1,204
Nova Scotia	228	1,085	1,313
New Brunswick	191	1,125	1,316
Québec	590	3,000	3,590
Ontario	1,392	4,307	5,699
Manitoba	205	1,205	1,410
Saskatchewan	213	1,168	1,381
Alberta	221	1,544	1,765
British Columbia	431	1,723	2,154
Total	3,778	17,276	21,054

5.2 1994–1995 Sample Design

The redesigned LFS was used as the basis for the design in all provinces except Québec, where the NPHS selected a sample from households already being used by Santé Québec for the 1992–1993 *Enquête sociale et de santé* (ESS).

5.2.1 Sample Design for the Household Component

Three factors shaped the design of the household component sample:

- the targeted national and provincial/territorial sample sizes;
- the decision to select one member per household to make up the longitudinal panel;
- the choice of the redesigned LFS as a vehicle for selecting the sample.

These three factors resulted, respectively, in the allocation of the sample, the application of a technique (the "rejective method," described later) to improve the sample's representativeness and the selection of provincial samples outside Québec.

5.2.2 Sample Allocation

The NPHS was budgeted for a sample size of 19,600 households. It was further agreed among national and provincial representatives that each province needed a minimum of 1,200 households. Subject to this restriction the provincial sample sizes were obtained by using a well-known allocation scheme that balances the reliability requirements at national and regional levels (Kish,

1988)[2]. According to this scheme, the sample was allocated proportionally to $\sqrt{(0.804Wh^2 + 1/12^2)}$, where W_h is the 1991 Census proportion of households in province/territory h=1,..,12. This allocation determined the base sample size for each province. Four provinces chose to increase their allotted sample size through the buy-in of additional units.

Within provinces, the sample was initially distributed proportionally to the population size. The provincial buy-in samples and the use of a rejective method, described below, affected the sub-provincial allocations. Ontario and Manitoba's buy-in samples imposed minimum requirements by health areas, while New Brunswick and British Columbia paid for additional sample coverage of certain areas only. In B.C. most of the buy-in requirement was met using telephone interviews from a Random Digit Dialling (RDD) sample of telephone numbers. In applying the rejective method, sample sizes were inflated by the number of households expected to be screened out of the sample. The «effective sample size» referred to earlier in this chapter excludes these screened-out households, as well as other out-of- scope dwellings, such as those under construction.

5.2.3 The Rejective Approach

The survey content primarily focused on one member in each sample household who was chosen at random to become the longitudinal panel respondent. Without the use of the rejective method, the panel would under-represent persons coming from large households, typically parents and children, since they had less chance of being chosen, and over-represent persons coming from small households, often single people or the elderly.

Thus, a rejective approach was adopted to increase the representation of parents and youths in the panel. A portion of the sample was pre-identified for screening. After their member roster was completed, screened households that had no member under 25 years of age were eligible for rejection and dropped from the survey. In order to maintain the required sample sizes, the number of households visited in each province was increased by the anticipated number of households to be screened out in this way.

The rejective method with an under-25-year-old rule was adopted as it performed better than other rejection rules considered. For cost and operational reasons, the percentage of screened households was usually limited to 25–30% in Ontario, 37.5–40% in urban areas elsewhere and 25–30% in rural areas. As apartment strata had a high concentration of small households, their sample sizes were reduced instead of applying a rejective method. The rejective approach was also not applied in remote regions because of the high contact

[2] Kish, L. (1988). Multipurpose Sample Design, *Survey Methodology*, 14, 19–32.

costs there, and its use was limited in areas where sample buy-in demands were substantial.

5.2.4 Sample Selection

The sample design considered for the household component of the NPHS was a stratified two-stage design. In the first stage, homogeneous strata were formed and independent samples of clusters were drawn from each stratum. In the second stage, dwelling lists were prepared for each cluster and dwellings were selected from the lists.

In all provinces except Québec, the NPHS used the multipurpose sampling methodology developed for the redesign of the LFS. That methodology provided general household surveys with clustered samples of dwellings, thus making the design very cost effective for the listing and collection of data.

The basic LFS design is a multistage stratified sample of dwellings selected within clusters. Each province is divided into three types of areas (Major Urban Centres, Urban Towns and Rural Areas) from which separate geographic and/or socio-economic strata are formed. In most strata, six clusters, usually Census Enumeration Areas (EAs), are selected with probability proportional to size (PPS). In a few cases where the population density was low, an additional stage is added by first selecting two or three large Primary Sampling Units, dividing them into clusters, and drawing a sample of six clusters from each. The number six is used throughout the sample design to allow a one-sixth rotation of the sample every month for the LFS.

The sample of dwellings is obtained once listing operations in sample clusters are completed. As sampling rates are predetermined, there are often differences between anticipated and obtained sample counts. Excessive sample yields are corrected by dropping a portion of the originally selected units. This is usually done at aggregated levels and is called sample stabilisation. Note also that sample sizes are inflated to represent dwellings rather than households, as approximately 15% of the dwellings are expected to be vacant or otherwise out of scope.

The LFS sample design is set up to yield about 60,000 households. Surveys needing smaller sample sizes usually "reserve" from one to six rotation groups per province, a rotation group being one-sixth of the total sample. Sample stabilisation is used to maintain the sample at desired levels, as when two rotation groups are reserved but the sample size needed only represents 1.5 rotation groups.

Requirements specific to the NPHS led to two modifications to this sampling strategy. The number of "reserved" rotation groups needed was specified at the stratum level rather than the provincial level in order to meet the specific sub-provincial sample size requirements. It was also required that the number of

clusters selected per stratum be a multiple of four for variance estimation and seasonal representativity (allowing each stratum to have two or more independent samples of four clusters each one per collection period). As the NPHS usually requested only between two and six clusters per LFS stratum, similar LFS strata were grouped to form larger NPHS strata with the required number of sample clusters.

As a result of these modifications, the NPHS sample of clusters can be considered as a stratified replicated sample, where strata are groups of LFS strata and replicates are typically independent, identically distributed samples of four clusters each. There were exceptions, but they are not expected to have a significant impact on survey results.

5.2.5 Sample Design in Québec

In Québec, the NPHS sample is selected from dwellings participating in a health survey organized by Santé Québec: the 1992–1993 *Enquête sociale et de santé* (ESS). The survey sampled 16,010 dwellings using a two-stage design similar to that of the LFS. The province was divided geographically by crossing fifteen health areas with four urban density classes (the Montreal Census Metropolitan Area, regional capitals, small urban agglomerations, and the rural sector). In each area, clusters were stratified by socio-economic characteristics and selected using a PPS sample. Selected clusters were enumerated and random samples of their dwellings were drawn: ten per cluster in major cities, twenty or thirty elsewhere.

Santé Québec provided non confidential information, which allowed the classification of its sample into four types of households: one-member households, households with children, other households with youths (persons aged under 25), and the rest (more than one member and no youth or child). NPHS personnel determined a household type for the ESS non-respondents.

The NPHS sample size was first allocated among the four urban density classes. To avoid having too much sample in Montreal, the allocation was proportional to $\sqrt{(2Wh^2 + 1/4^2)}$, where W_h is the population share for class h=1,2,3,4. In each class, an attempt was made to obtain a sub-sample from the ESS that, as far as the selected panel member was concerned, would be proportional to the populations for the four household types. This was done by drawing a sufficient number of households from the ESS to give the required yield for households with children (the most underrepresented group), and then removing excess sample from the other three household groups. An initial sample that was almost 50% higher than needed was thus selected. After removing from it two-thirds of the one-member households, one-half of the other households with no youths or children, and one-sixth of households with youths but no children, the objective was largely attained.

Considerations for seasonal representation and variance estimation, and integration with the NLSCY, affected the sub-sampling in Québec as they did elsewhere. ESS strata were thus collapsed to allow the formation of replicates, with the clusters in each replicate covering all four quarters (two quarters are covered per cluster in the rural and small urban sectors due to their larger sample sizes). The sample of households with children was split into an "Adult" sample and a "Child" sample by a 3:2 ratio, the terms having the same meaning as in other provinces. "Child" sampled households in quarters 1 and 2 were reassigned to collection in quarters 3 and 4. Since the NPHS surveyed the current occupants of dwellings selected for the ESS, and changes occurred in some of those dwellings, the samples of households without children for quarters 3 and 4 were also to be split, by a 2:3 ratio, into an "Adult" and a "Child" sample.

6. Data Collection

6.1 Questionnaire Design and Data Collection Method

The NPHS questions were designed for computer-assisted interviewing (CAI), meaning that, as the questions were developed, the associated logical flow into and out of the questions was programmed. This included specifying the type of answer required, the minimum and maximum values, on-line edits associated with the question and what to do in case of item non-response.

With CAI, the interview can be controlled based on answers provided by the respondent. On-screen prompts are shown when an invalid entry is recorded and thus immediate feedback is given to the respondent and/or the interviewer to correct inconsistencies. Another enhancement is the automatic insertion of reference periods based on current dates. Pre-filling of text or data based on information gathered during the interview allows the interviewer to proceed without having to search back for previous answers. This type of pre-fill includes such things as using the correct name or sex within the questions themselves. Allowable ranges/answers based on data collected during the interview can also be programmed. In other words, the questionnaire can be customized to the respondent based on data collected at that time or during a previous interview.

6.2 Tests

Two field tests were conducted. The tests involved four of Statistics Canada's Regional Offices. Experienced Labour Force Survey interviewers carried out interviews. The main objectives of the two tests were to observe respondent reaction to the survey, to obtain estimates of time for the various sections, to study the response rates and to test feedback questions. Field operations and procedures, interviewer training and the data collection computer application were also tested.

In addition to the field tests, the data collection computer application was extensively tested in-house in order to identify any errors in the program flow and text. The testing of the data collection computer application was an ongoing operation up until the start of the main survey.

6.3 Interviewing

The interviewers were trained specifically to carry out surveys using the computer-assisted interviewing method. Collection was divided into four quarters (June, August and November 1998 and February 1999). An additional collection was held in June 1999 with further tracing attempts of non-respondents from previous quarters. The LFS supervisory and control structure was employed for the NPHS collection.

Respondents in the sample and the top-up sample of households with young children were first contacted by telephone. 95% of the interviews were done by telephone.

Personal visits were made if the respondent did not have a telephone, if the interviewer made a personal visit in the course of tracing a respondent or upon request by the respondent. Respondents in the top-up sample of cycle 1 non-respondents and the top-up sample of households with recent immigrants were also first contacted with a personal visit by the interviewer. The total interview took an average of one hour in each household.

In all dwellings, information about all household members was obtained from the person at home at the time of the interviewer call. However in 1998 we encouraged the interviewers to start the interviews with the longitudinal respondents. Approximately 15% of the information collected for this part of the interview (for longitudinal respondents, 12 years and over) was done by proxy.

Proxy reporting of the health component was allowed for the longitudinal respondent only for reasons of illness or incapacity. Such proxy reporting accounted for 2.7% of the information collected for respondents aged 12 years and older. On the other hand, all interviews for longitudinal respondents under 12 years old were done by proxy.

6.4 Non-response and Tracing

Interviewers were instructed to make all reasonable attempts to obtain NPHS interviews with members of eligible households. For individuals who at first refused to participate in the NPHS, a letter was sent from the 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 the interviewer. For cases in which 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.

Many strategies were put in place to reduce the number of non-response cases. Before interviews started, the maximum recommended assignment size by interviewer was calculated based on test results. This allowed for efficient follow-up of non-contact cases (i.e., avoid overburdening interviewers). Interviewer training covered ways of reducing the number of non-contacts (e.g., making calls or visits at various times of the day) using contact information given in the previous interview.

Refusals were followed up by senior interviewers, project supervisors or by other interviewers to try to convince respondents of the importance of participating in the survey. To maximize the response rate, a large number of non-response cases were also followed up in subsequent collection periods.

The failure to trace a longitudinal respondent was an additional type of non-response. Interviewers had several ways to trace a respondent. The last known address and telephone number were provided as part of the information on the case, as well as the name and address of one or two previous contacts, if collected. In addition, interviewers were trained to follow up available leads such as local telephone directories and directory assistance. If these leads were unsuccessful, the case was transmitted to an experienced interviewer specially trained in tracing respondents.

Tracer interviewers had access to Canada-wide telephone directories and reverse directories. The cumulative non-response rate due to failure to trace the longitudinal respondent is 2.1%, which is still exceedingly low.

7. Data Processing

7.1 Editing

Editing was performed on-line in the computer-assisted interviewing (CAI) application during data collection. 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 other situations, warning messages were invoked, but no corrective action was taken (e.g., if an interviewer entered contradictory responses between questions). Because no corrective action was taken in such instances, edits were 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". No imputation was performed.

7.2 Coding

Several questions allowing write-in responses had the write-in information coded into either new unique categories or to an existing listed category if the write-in information duplicated a listed category. Where possible (e.g., occupation, industry, diseases), the coding followed the standard classification systems as used either in the Census of Population or in other Statistics Canada surveys such as the Health and Activity Limitation Survey and General Social Survey-cycle 6.

7.3 Creation of Derived and Grouped Variables

To facilitate data analysis, a number of variables on the file have been derived using items found on the NPHS questionnaires. Derived variables generally have a "D" or "G" 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. Appendix F provides details on how these more complex variables were derived.

7.4 Weighting

The principle behind estimation in a probability sample such as the NPHS 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 (see question SMC8_2 in section 9.2) 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 Chapter 11.

7.5 Suppression of Confidential Information

It should be noted that the "public use" microdata files differ in a number of important aspects from the survey "master" files held by Statistics Canada. These differences are the result of actions taken to protect the anonymity of individual survey respondents through suppression of individual values, variable grouping, and variable capping. Statistics Canada offers various options that allow the calculation of statistics, which includes the information excluded from the microdata files. These options are described in Section 12.3. Estimates generated by remote access will be released to the user, subject to meeting the guidelines for analysis and release outlined in Chapter 9 of this document.

8. Data Quality

8.1 Response Rates

The calculation of response rates for the NPHS must take into account the augmentation of the sample by returning to cycle 1 non-responding dwellings and by the LFS top-up sample (composed of immigrants and children), used for cross-sectional estimates only. Cross-sectional response rates are thus calculated separately for the core, for the cycle 1 non-respondent dwellings and for the LFS top-up, and overall. The following table contains a summary of the 1998-1999 response rates:

1998-1999 Response Rates Table

Level	Core	Cycle 1 non-respondent dwellings	LFS top-up	Overall
Household	89.7%	71.5%	92.5%	87.6%
Selected-person	98.8%	96.4%	98.8%	98.5%

The following is a description of how the household response rate and the selected-person response rates were calculated, for the cross-sectional file. Note that in the four sections below, the selected-person response rate is calculated based on the number of responding households. To get an idea of the overall rate of response for selected person, the two rates can be multiplied.

8.1.1 Core Cross-sectional Response Rates

Household (HH) response rate

$$\frac{\textit{\# of responding households}}{\textit{all in-scope continuing households}}$$

The 1998-1999 *core cross-sectional* response rate is based on all continuing households. A responding household had *at least* one general-component questionnaire (H05) completed for a member of the household. This response rate at the Canada level for the NPHS was ***89.7%***. At the provincial level, this rate varied from 85.3% in British Columbia to 93.1% in Newfoundland.

Selected-person (SP) response rate

The core selected-person response rate can be thought of as the number of health-component questionnaires (H06) that *were* completed compared to the number of health-components that *should have been* completed.

$$\frac{\textit{\# of responding H06s}}{\textit{\# of responding households (i.e., eligible to answer)}}$$

For the core, the selected-person response rate for the NPHS was ***98.8%*** at the Canada level, and ranged from 98.2% in Québec to 99.6% in Saskatchewan.

Relevant information for calculation of response rates:

Number of continuing households:	17,276
Number of out-of-scope households:	880
Number of respondents at the household level:	14,702
Number of respondents at the selected-person level:	14,520
Number of non-respondents at the household level:	1,694
Number of non-respondents at the selected-person level:	182

Calculation of household response rate:

$$\text{HH Rate} = \frac{14{,}702}{17{,}276 - 880} = \frac{14{,}702}{16{,}396} = 89.7\%$$

Calculation of selected-person response rate:

$$\text{SP Rate} = \frac{14{,}520}{14{,}702} = 98.8\%$$

8.1.2 Cycle 1 Non-responding Dwelling Sample Cross-sectional Response Rates

Household (HH) response rate

$$\frac{\textit{\# of responding households}}{\textit{\# of household living in non-responding in – scope cycle 1 dwellings}}$$

The 1998-1999 cross-sectional response rate for the household living in non-responding Cycle 1 dwellings is based on a sample of the non-responding dwellings from the cycle 1 core sample, except particular out-of-scope

dwellings (status code = 013, 017, 023, 024)[3]. A responding household had *at least* one general-component questionnaire completed for a member of the household. This response rate at the Canada level for the NPHS was ***71.5%***. At the provincial level, this rate varied from 59.5% in Prince Edward Island to 76.5% in Ontario.

Selected-person (SP) response rate

The selected-person response rate can be thought of as the number of health-component questionnaires that *were* completed compared to the number of health-components that *should have been* completed.

$$\frac{\#\ of\ responding\ H06s}{\#\ of\ responding\ households\ (i.e.,\ eligible\ to\ answer)}$$

For the cycle 1 non-respondents portion, the selected-person response rate for the NPHS was ***96.4%*** at the Canada level, and ranged from 94.4% in British Columbia to 100.0% in Newfoundland.

Relevant information for calculation of response rates:

Number of sampled non-responding dwellings from the cycle 1 core sample:	2,432
Number of out-of-scope households:	8
Number of respondents at the household level:	1,732
Number of respondents at the selected-person level:	1,670
Number of non-respondents at the household level:	692
Number of non-respondents at the selected-person level:	62

Calculation of household response rate:

$$\text{HH Rate} = \frac{1,732}{2,432 - 8} = \frac{1,732}{2,424} = 71.5\%$$

Calculation of selected-person response rate:

$$\text{SP Rate} = \frac{1,670}{1,732} = 96.4\%.$$

[3] 013=Business, institution or other non-residence.
017=Other ineligible dwelling (e.g., embassy).
023=Under construction or demolished.
024=Dwelling vacant.

8.1.3 LFS Top-up Sample Cross-sectional Response Rates

Household (HH) response rate

$$\frac{\textit{\# of responding households}}{\textit{\# of in-scope households selected from the LFS}}$$

The 1998-1999 *LFS top-up sample cross-sectional* response rate is based on all the households selected from the LFS. A responding household had *at least* one general-component questionnaire completed for a member of the household. This response rate at the Canada level for the NPHS was ***92.5%*** (***92.5%*** for infants and ***92.7%*** for new immigrants)**.** At the provincial level, this rate varied from 89.6% in New Brunswick to 97.3% in Saskatchewan.

Selected-person (SP) response rate

The selected-person response rate can be thought of as the number of health-component questionnaires that *were* completed compared to the number of health-components that s*hould have been* completed.

$$\frac{\textit{\# of responding H06s}}{\textit{\# of responding households (i.e., eligible to answer)}}$$

For the LFS top-up portion, the selected-person response rate for the NPHS was ***98.8%*** (***100.0%*** for infants and ***96.5%*** for new immigrants) at the Canada level, and ranged from 97.6% in Alberta to 100.0% in Newfoundland, Prince Edward Island, Nova Scotia, New Brunswick, Québec and Saskatchewan.

Relevant information for calculation of response rates:

Number of households selected from the LFS frame:	1,179
- Infants:	758
- New immigrants:	421
Number of out-of-scope households:	26
Number of respondents at the household level:	1,067
Number of respondents at the selected-person level:	1,054
Number of non-respondents at the household level:	86
Number of non-respondents at the selected-person level:	13

Calculation of household response rate:

$$\text{HH Rate} = \frac{1{,}067}{1{,}179 - 26} = \frac{1{,}067}{1{,}153} = 92.5\%$$

Calculation of selected-person response rate:

$$\text{SP Rate} = \frac{1{,}054}{1{,}067} = 98.8\%$$

8.1.4 Overall Cross-sectional Response Rates

Household (HH) response rate

$$\frac{\textit{\# of responding households}}{\textit{all in-scope NPHS cycle 3 selected households}}$$

For the core, the cycle 1 non-respondent sample and the LFS top-up sample households combined, the cross-sectional response rate at the Canada level was ***87.6%***. At the provincial level, this rate varied from 83.6% in British Columbia to 91.0% in Newfoundland.

Selected-person (SP) response rate

The selected-person response rate can be thought of as the number of health-component questionnaires that *were* completed compared to the number of health-components that *should have been* completed.

$$\frac{\textit{\# of responding H06s}}{\textit{\# of responding households (i.e., eligible to answer)}}$$

The selected-person response rate was ***98.5%*** at the Canada level, and ranged from 98.2% in Nova Scotia and Québec to 99.3% in Saskatchewan.

Relevant information for calculation of response rates:

Number of NPHS cycle 3 selected households:	20,887
Number of out-of-scope households:	914
Number of respondents at the household level:	17,501
Number of respondents at the selected-person level:	17,244
Number of non-respondents at the household level:	2,472
Number of non-respondents at the selected-person level:	257

Calculation of household response rate:

$$\text{HH Rate} = \frac{17{,}501}{20{,}887 - 914} = \frac{17{,}501}{19{,}973} = 87.6\%$$

Calculation of selected-person response rate:

$$\text{SP Rate} = \frac{17{,}244}{17{,}501} = 98.5\%$$

8.2 Survey Errors

The survey produces estimates based on information collected from a sample of individuals. Somewhat different estimates might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. as those used in the survey. The difference between the estimates obtained from the sample and those resulting from a complete count taken under similar conditions is called the sampling error of the estimate.

Errors that 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 and errors may be introduced in the processing and tabulation of the data. These are all examples of 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 survey. Quality assurance measures were implemented at each step of the data collection and processing cycle to monitor the quality of the data. These measures included the use of highly-skilled interviewers, extensive training of interviewers with respect to the survey procedures and questionnaire, observation of interviewers to detect problems and 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 NPHS was basically non-existent; once the questionnaire was started, it tended to be completed with very little non-response. Total non-response occurred because the interviewer was either unable to trace the respondent, no member of the household was able to provide the information or the respondent refused to participate in the survey. Total non-response was handled by adjusting the weight of households that responded to the survey to compensate for those who did not respond.

In most cases, partial non-response to the survey occurred when the respondent did not understand or misinterpreted a question, refused to answer a question, could not recall the requested information or could not provide personal or proxy information.

This section of the documentation outlines the measures of sampling error that Statistics Canada commonly uses and that it urges users producing estimates from this microdata file to use also. 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 error of the estimates derived from survey results.

However, because of the large variety of estimates that can be produced from a survey, the standard error of an estimate is usually expressed relative to the estimate to which it pertains. This resulting measure, known as the coefficient of variation (C.V.) of an estimate, is obtained by dividing the standard error of the estimate by the estimate itself and is expressed as a percentage of the estimate.

For example, suppose that based upon the survey results, one estimates that 24% of Canadians aged 12 and over are daily cigarettes smokers is found to have standard error of .003. Then the coefficient of variation of the estimate is calculated as:

$$\left(\frac{.003}{.24}\right) x\, 100\% = 1.25\%$$

9. 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 files. 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.

9.1 Rounding Guidelines

In order that estimates for publication or other release derived from these microdata files 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.

9.2 Sample Weighting Guidelines for Tabulation

The sample design used for the NPHS 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.

9.2.1 Definitions: Categorical Estimates, Quantitative Estimates

Before discussing how the NPHS 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 for the National Population Health Survey.

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:

SMK-Q2: At the present do/does ... smoke cigarettes daily, occasionally or not at all?

__ 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:

SMK-Q4: How many cigarettes do/does you/he/she smoke each day now?

|_|_| Number of cigarettes

9.2.2 Tabulation of Categorical Estimates

Estimates of the number of people with a certain characteristic can be obtained from the microdata files y 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.

9.2.3 Tabulation of Quantitative Estimates

Estimates of quantities can be obtained from the microdata files by:

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 an estimate of the average number of cigarettes smoked each day by individuals who smoke daily, multiply the value of variable *SMC8_4*[4] (question SMK-Q4) by the weight, WT68, then sum this value over those records with a value of "daily" to the variable SMC8_2 (question SMK-Q2) to obtain the numerator ($\hat{X}$). Sum the final weight of those records with a value of "daily" to the variable SMC8_2 (question SMK-Q2) to obtain the denominator ($\hat{Y}$). Divide ($\hat{X}$) by ($\hat{Y}$) to obtain the average number of cigarettes smoked each day by daily smokers.

Note: See Section 12.2 for variable naming convention

9.3 Guidelines for Statistical Analysis

The National Population Health Survey 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 differs 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 "C.V. Tables") for the NPHS. These tables can be used to obtain approximate coefficients of variation for categorical-type estimates and proportions. See Chapter 10 for more details.

[4] See Section 12.2 for variable naming convention

9.4 Release Guidelines

Before releasing and/or publishing any estimate from these microdata files, users should first determine the number of sampled respondents who contribute to the calculation of the estimate. 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 30 or more, users should determine the coefficient of variation of the rounded estimate and follow the guidelines below.

Sampling Variability Guidelines

Type of Estimate	C.V. (in %)	Guidelines
1. Acceptable	0.0 - 16.5	Estimates can be considered for general unrestricted release. Requires no special notation.
2. Marginal	16.6 - 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 M (or in some other similar fashion).
3. Unacceptable	Greater than 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 U (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.»

10. Approximate Coefficients of Variation Tables

In order to supply coefficients of variation that would be applicable to a wide variety of categorical estimates produced from these microdata files and that could be readily accessed by the user, a set of Approximate Coefficients of Variation Tables (C.V. 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 (C.V.) are derived using the variance formula for simple random sampling and incorporating a factor that 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 from among these a conservative value to be used in the look-up tables, which would then apply to the entire set of characteristics.

The four tables below show the design effects, sample sizes and population counts used to produce the four sets of Approximate Coefficients of Variations Tables. The four sets correspond to: the provincial and Canada levels for both household members and selected members and various age groups at the Canada level for both household members and selected members.

Input Data for Provincial and Canada Level
Approximate Coefficients of Variation Tables
For Household Members (All Ages)

Province	Design effect	Sample size	Population
Newfoundland	1.42	2,871	538,051
Prince Edward Island	1.46	2,755	135,015
Nova Scotia	1.42	2,977	910,115
New Brunswick	1.45	2,959	734,217
Québec	1.84	8,305	7,157,967
Ontario	1.55	13,526	11,261,692
Manitoba	1.88	3,359	1,069,459
Saskatchewan	1.40	3,092	971,067
Alberta	1.45	4,541	2,846,030
British Columbia	1.43	4,661	3,900,651
CANADA	1.94	49,046	29,524,264

Input Data for Canada Level Age Group Approximate Coefficients of Variations Tables For Household Members

Age Group	Design effect	Sample size	Population
0–11	2.55	9,698	4,608,137
12–24	2.36	9,126	5,203,074
25–44	2.14	15,590	9,548,249
45–64	1.78	9,904	6,677,191
65+	1.76	4,728	3,487,613

Input Data for Provincial and Canada Level Approximate Coefficients of Variations Tables For Selected Members (All Ages)

Province	Design effect	Sample size	Population
Newfoundland	0.98	963	538,051
Prince Edward Island	0.92	932	135,015
Nova Scotia	0.95	1,071	910,116
New Brunswick	1.01	1,073	734,217
Québec	1.14	2,946	7,157,967
Ontario	1.22	4,691	11,261,691
Manitoba	1.19	1,184	1,069,459
Saskatchewan	0.97	1,131	971,067
Alberta	1.03	1,568	2,846,030
British Columbia	1.01	1,685	3,900,651
CANADA	1.53	17,244	29,524,264

Input Data for Canada Level Age Group Approximate Coefficients of Variations Tables For Selected Members

Age Group	Design effect	Sample size	Population
0–11	1.65	1,995	4,608,137
12–24	1.67	2,526	5,203,074
25–44	1.77	5,775	9,548,249
45–64	1.76	4,097	6,677,191
65+	1.71	2,851	3,487,613

All coefficients of variation in the C.V tables are approximate and, therefore, unofficial. Estimates of actual variance for specific variables may be obtained from Statistics Canada. The use of actual variance estimates would allow users to release otherwise unreleaseable estimates, i.e., estimates with coefficients of variation in the "unacceptable" range. See section 10.7 for details.

Remember: 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 for this estimate. This is because the formulas used for estimating the variance do not hold true for small sample sizes.

10.1 How to Use the Approximate C. V. Tables for Categorical Estimates

The following rules should enable the user to determine the approximate coefficients of variation from the C.V. 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. 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 or percentage and the size of the total upon which the proportion or percentage is based. Estimated proportions or percentages are relatively more reliable than the corresponding estimates of the numerator of the proportion or percentage, when the proportion or percentage is based upon a subgroup 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 C.V.'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 C.V. of the proportion or percentage is the same as the C.V. of the numerator of the proportion or percentage. In this case, Rule 1 can be used.

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 or percentage (across the top of the table) and to the numerator of the proportion or percentage (down the left side of the table). The intersection of the appropriate row and column gives the coefficient of variation.

Rule 3: Estimates of Differences Between Aggregates or Percentages

The standard error of a difference between two estimates is approximately equal to the square root of the sum of squares of each standard error considered separately.

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

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

where $\hat{X}_1$ is estimate 1, $\hat{X}_2$ is estimate 2, and α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. The coefficient of variation of $\hat{d}$ is given by $\sigma_{\hat{d}} / \hat{d}$. This formula is accurate for the difference between separate and uncorrelated characteristics, but is only approximate otherwise.

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 C.V.'s for the two ratios are first determined using Rule 4, and then the C.V. of their difference is found using Rule 3.

10.2 Examples of Using the C.V. 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 5,690,625 individuals smoke daily in Canada. How does the user determine the coefficient of variation of this estimate?

1) Refer to the Canada level C.V. table for SELECTED MEMBERS;

2) The estimated aggregate (5,690,625) does not appear in the left-hand column (the "Numerator of Percentage" column), so it is necessary to use the figure closest to it, namely 6,000,000;

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, 1.8%.

4) So the approximate coefficient of variation of the estimate is 1.8%. The finding that there were 5,690,625 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 5,690,625/6,622,143=85.9% 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 C.V. table for SELECTED MEMBERS;

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 (85.9%) and the numerator portion of the percentage (5,690,625) in determining the coefficient of variation;

3) The numerator, 5,690,625, does not appear in the left-hand column (the "Numerator of Percentage" column) so it is necessary to use the figure closest to it, namely 6,000,000. Similarly, the percentage estimate does not appear as any of the column headings, so it is necessary to use the figure closest to it, 90.0%;

4) The figure at the intersection of the row and column used, namely 0.7% is the coefficient of variation (expressed as a percentage) to be used;

5) So the approximate coefficient of variation of the estimate is 0.7%. The finding that 85.9% 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 4,738,935/5,690,625=83% of those who smoke daily smoke 10 or more cigarettes daily (estimate 1) while 4,661,275/6,287,675=74% of those who smoke occasionally or not at all, but at one time smoked daily, smoked 10 or more cigarettes daily at that time (estimate 2). Note that these estimates are based on the value of the variables SMC8_2, SMC8_4, SMC8_4A, SMC8_5 and SMC8_7. How does the user determine the coefficient of variation of the difference between these two estimates?

1) Using the Canada level C.V. table for selected members in the same manner as described in example 2 gives the C.V. for estimate 1 as 0.7% (expressed as a percentage), and the C.V. for estimate 2 as 1.3% (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.

That is, the standard error of the difference $\hat{d} = (.74 - .83) = .09$ is:

$$\sigma_{\hat{d}} = \sqrt{[(.83)(.007)]^2 + [(.74)(.013)]^2}$$
$$= .011$$

3) The coefficient of variation of $\hat{d}$ is given by $\sigma_{\hat{d}} / \hat{d} = .011/.09 = 0.122$;

4) So the approximate coefficient of variation of the difference between the estimates is 12.2% (expressed as a percentage). This estimate can be published with no qualifications;

Example 4: Estimates of Ratios

Suppose that the user estimates that 5,690,625 individuals smoke daily, while 931,518 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 C.V. table for selected members;

3) The numerator of this ratio estimate is 931,518. The 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, 5.0%;

4) The denominator of this ratio estimate is 5,690,625. The figure closest to it is 6,000,000. 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, 1.8%;

5) So the approximate coefficient of variation of the ratio estimate is given by rule 4, which is, $\alpha_{\hat{R}} = \sqrt{\alpha_1^2 + \alpha_2^2}$ where α_1 and α_2 are the coefficients of variation of $\hat{X}_1$ and $\hat{X}_2$ respectively. That is

$$\alpha_{\hat{R}} = \sqrt{(.050)^2 + (.018)^2} = 0.053$$

The obtained ratio of occasional to daily smokers is 931,518/5,690,625 is 0.16:1. The coefficient of variation of this estimate is 5.3% (expressed as a percentage), which is releasable with no qualifications.

10.3 How to Use the C.V. 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 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 Coefficients of Variation 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 the determined coefficient of variation of $\hat{X}$, 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, 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.

10.4 Example of Using the C.V. 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, section 10.2) would be calculated as follows.

$\hat{X}$ = .859

z = 2

$\alpha_{\hat{X}}$ = .007 is the coefficient of variation of this estimate as determined from the tables

CI_X = {.859 - (2) (.859) (.007), .859 + (2) (.859) (.007)}

CI_X = {.847, .871}

10.5 How to Use the C.V. 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 X_1 and 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 $z = (\hat{X}_1 - \hat{X}_2) / \sigma_{\hat{d}}$ 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.

10.6 Example of Using the C.V. Tables to do a Z-test

Let us suppose one wants to test, at 5% level of significance, the hypothesis that there is no difference between the proportion of individuals who smoke daily at a rate of 10 or more cigarettes AND the proportion of those who smoke occasionally or not at all, but at one time smoked daily at a rate of 10 or more cigarettes. From example 3, section 10.2, the standard error of the difference between these two estimates was found to be = .011. Hence ,

$$z = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_{\hat{d}}} = \frac{.83 - .74}{.011} = \frac{.09}{.011} = 8.18$$

Since z = 8.18 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.

10.7 Exact Variances/Coefficients of Variation

All coefficients of variation in the C.V. Tables are approximate and, therefore, unofficial. Statistics Canada can provide exact coefficients of variation for specific estimates (see Section 12.3 for more details). The types of estimates supported include aggregates, proportions, ratios, and differences between aggregates, as well as more sophisticated types of analyses such as estimates of coefficients from linear regressions and logistic regressions, among others. The exact coefficients of variation are obtained via an exact variance program, which uses a technique called "bootstrapping". This technique involves dividing the records on the microdata files into subgroups (or

replicates) and determining the variation in the estimates from replicate to replicate. 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 smaller than the province (for example, at the urban/rural level), then the C.V. 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 coefficients from linear regressions or logistic regressions, the C.V. 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 National Population Health Survey 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.

Finally, should a user find himself/herself in a position where he/she can use the C.V. 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.

Users can obtain the exact variance/coefficient of variation program or submit custom requests by contacting the Client Custom Services, Health Statistics Division, Statistics Canada at **1-613-951-1746** or by e-mail at **hd-ds@statcan.ca**. There will be a fee charged for any consultation time required to set up the request as well as for any time required to set up the associated computer runs. Also, the public use microdata files users can have recourse, free of charge, to the Remote Access Services by sending their

duly tested programs by e-mail at **nphs@statcan.ca** or by contacting Colette Koeune, Health Statistics Division, Statistics Canada at **1-613-951-1653**.

10.8 Release Cut-offs for the NPHS

The minimum cut-offs for estimates of totals at the provincial and Canada levels and those for various age groups at the Canada level, for both household members and selected members, are specified in the four tables below. Statistics Canada recommends not to release estimates for which the estimate size is smaller than the minimum given in the "Marginal" column. However, if the user chooses to do so then he should indicate that the quality of the estimates is unacceptable and flag them with the letter U (or in some other fashion) with a warning (see section 9.4).

Table of Release Cut-offs for Totals Based on Provincial/Canada Level Estimates for Household Members (All Ages)

PROVINCE	ACCEPTABLE C.V. from 0% to 16.5%	MARGINAL C.V. from 16,6% to 33,3%
Newfoundland	9,500	2,500
Prince Edward Island	2,500	500
Nova Scotia	15,500	4,000
New Brunswick	13,000	3,000
Québec	58,000	14,500
Ontario	47,000	11,500
Manitoba	21,500	5,500
Saskatchewan	16,000	4,000
Alberta	33,000	8,000
British Columbia	43,500	11,000
CANADA	43,000	10,500

Table of Release Cut-offs for Totals Based on Age Group Estimates at the Canada Level for Household Members

AGE GROUP	ACCEPTABLE C.V. from 0% to 16.5%	MARGINAL C.V. from 16,6% to 33,3%
0–11	44,000	11,000
12–24	49,000	12,000
25–44	48,000	12,000
45–64	44,000	11,000
65+	47,000	11,500

Table of Release Cut-offs for Totals Based on Provincial/Canada Level Estimates for Selected Members (All Ages)

PROVINCE	ACCEPTABLE C.V. from 0% to 16.5%	MARGINAL C.V. from 16,6% to 33,3%
Newfoundland	19,500	5,000
Prince Edward Island	4,500	1,000
Nova Scotia	28,500	7,000
New Brunswick	24,500	6,000
Québec	100,500	25,000
Ontario	106,500	26,500
Manitoba	38,000	9,500
Saskatchewan	29,500	7,500
Alberta	67,000	17,000
British Columbia	84,000	21,000
CANADA	96,000	23,500

Table of Release Cut-offs for Totals Based on Age Group Estimates at the Canada Level for Selected Members

AGE GROUP	ACCEPTABLE C.V. from 0% to 16.5%	MARGINAL C.V. from 16,6% to 33,3%
0–11	136,000	34,000
12–24	123,500	31,000
25–44	106,500	26,500
45–64	103,500	26,000
65+	75,000	19,000

11. Weighting

The household component of the National Population Health Survey in 1994-1995 had two basic designs, one for nine of the provinces, and one for Québec. In the nine provinces other than Québec, the NPHS used the design of the Labour Force Survey (LFS), with many modifications, to generate a sample of its own. Consequently, the derivation of weights was tied to the weighting procedure used for the LFS. In Québec, however, a two-phase sample design was implemented, where the first phase was drawn by the *Enquête sociale et de santé* (ESS) in 1992–1993, and the second phase sample was drawn by the NPHS. Thus, in Québec, the derivation of the weights was tied to the weighting procedure used by the ESS. See Chapter 5, Sample Design for more details. In 1996–1997, additional independent samples were drawn in Ontario, Manitoba and Alberta using Random Digit Dialing (RDD) to allow for the production of reliable estimates at the health-area level for cycle 2 only. In 1998–1999, a supplemental sample was drawn to make the overall cross-sectional sample more representative of the population and thus, improve the cross-sectional estimates. This better representation was obtained by first compensating for sample attrition and secondly by accounting for initially absent persons. The former was achieved by including all nonresponding dwellings from cycle 1, while the latter was achieved by including a LFS subsample of households that contained infants or recent immigrants, two subpopulations absent from the original sample in 1994-1995.

In cycle 1, two sets of weights were required, one for cross-sectional estimates based upon variables from the general component of the questionnaire administered in 1994–1995 to all household members, and one for cross-sectional estimates based on variables from the health component of the questionnaire administered to only the selected member. In cycle 2, three sets of weight were required for cross-sectional purposes. The first two were similar to those of 1994–1995, but were calculated differently to take into account the sample design of 1996–1997. The third weight was applicable only to the health file and was required for analysis of specific populations and variables. In cycle 3, two sets of weights only are required. These are once again similar to those of 1994–1995, but calculated differently to take into account the 1998–1999 sample design.

The 1998–1999 weighting procedure is based significantly on that of 1994–1995. The 1994–1995 final weights for each continuing selected respondent are taken as a starting point. Some weight adjustments that may be different for 1998–1999 have been removed, to create a «stripped» weight for each selected respondent. From this point, the new adjustments are made to come up with the 1998–1999 final weight, for both the selected (longitudinal) member and every member of the current household. In the following sections, a brief description of the 1994–1995 weighting procedures still relevant for weighting in 1998–1999 is included. A full description of the 1994–1995 procedures may be found in the 1994–1995 National Population Health Survey Public-Use Microdata Files documentation. Section 11.1 describes the new adjustments necessary in 1998–1999 for the continuing (core) NPHS sample, while Section 11.2 describes the 1998–1999 procedures for the additional cross-sectional top-up sample. Section 11.3 and Section 11.4 describe the 1994–1995-based procedures for the provinces other than Québec and for Québec, respectively.

11.1 Cross-sectional Weighting for the NPHS Cycle 3, Core Household Sample

This section describes the 1998–1999 weighting procedures for selected members and all members of their households in the continuing (core) NPHS sample. A complete description of the additional weighting procedures necessary for the supplemental top-up sample is found in Section 11.2.

11.1.1 Stripped Weights

As described in Sections 11.3 and 11.4, the starting point of the 1998–1999 weighting procedure is the «stripped» weight, based upon the original sample design of 1994–1995. Essentially, the «stripped» weights are the inverse basic probabilities of selection for the dwellings. Once these weights are obtained, weight adjustments are performed as described in sections 11.1.2.1 and 11.1.2.2 for all provinces except for Québec. Adjustments for Québec are described in sections 11.1.2.3 and 11.1.2.4.

11.1.2 Adjustments to the Stripped Weights

11.1.2.1 Adjustments for all Provinces Other Than Québec

Adjustment 4: Multiples Weight Adjustment

It sometimes happens that an interviewer discovers that a listing entry thought to constitute single private occupied dwelling in fact constitutes two or more private occupied dwellings. This may happen, for instance, when a basement apartment is attached to a dwelling but has its own separate entrance. In this case, since interviewing takes place in only one of the private occupied dwellings (selected at random), the weight associated with that dwelling is boosted. Thus, the fourth multiplicative weight adjustment is given by the number of private occupied dwellings that the listing entry in question actually constitutes. For most listings, this adjustment factor will be one.

Note that this adjustment had to be recalculated to take into account the handful of multiples found in the top-up sample of cycle 1 nonresponding dwellings.

Adjustment 5: Cycle 1 Household Nonresponse Weight Adjustment

Despite all the attempts made by the interviewers, some nonresponse at the household level is inevitable. Nonresponse encompasses any of the following situations: refusal, special circumstance, language barrier, no one at home, temporarily absent or computer problem.

Nonresponse is compensated for by proportionally adjusting the weights of responding households.

It is also at this step that the removal of 1994–1995 cross-sectional buy-in units, which were not to be surveyed in 1998–1999, was accounted for. Households selected in 1994–1995 as part of the buy-in are removed from calculations. This fifth adjustment is done separately for the panel and the top-up sample. For the panel sample, the adjustment is given by the following:

$$\frac{\textit{sum of weights for sampled households in NPHS stratum/season combination}}{\textit{sum of weights for responding households in NPHS stratum/season combination}}.$$

This adjustment is made at the NPHS stratum level for each «season», i.e., summer and winter. Here, NPHS strata are groups of LFS strata. The adjustment was made at this level since it was the smallest geographic level that ensured stability. The adjustment was calculated separately for each season since the nonresponse rate was significantly different for each season. The "weights" referred to above are the LFS basic weight multiplied by all the adjustments to this point (i.e., weight adjustments 1 through 4). The adjustment is based on the assumption that the households that were actually interviewed represent the characteristics of those that were not interviewed in each stratum/season combination.

For the top-up sample, adjustments were done separately for the nonresponding cycle 1 dwellings and for the remaining part of the top-up, i.e. the infants and recent immigrants. The cycle 3 household nonresponse is dealt with here, in a similar fashion to what was done for the panel sample. However, because of small sample sizes, the adjustment was done at the province/CMA/urban level. Note that all out-of-scope dwellings (under construction, rejected households, etc.) are dropped.

Adjustment 6: Rejective Method Weight Adjustment

As discussed in Chapter 5, Sample Design, in the last two quarters of data collection in 1994–1995 a portion of the sampled households was screened out or rejected from the sample after determining that there were no youths or children residing within (i.e., no one less than the age of 25). These "rejected" households come from that portion of the "Children Sample" that are "screened" for household composition.

This methodology was implemented to compensate for an over-representation in the sample of members of small households and an

under-representation of members of large households, while maintaining the desired overrepresentation of elderly individuals. The large households often consist of parents and their children while the small ones tend to consist of single people, older people or couples without children. Since some households containing no youths or children are screened out or "rejected", representation in the sample of households of this type comes solely from the "Adult Sample" and from the non-screened portion of the "Children Sample". Thus, to compensate for the "rejected" part of the sample, the weights for those third and fourth quarter households containing no youths or children from the "Adult Sample" and from the non-screened portion of the "Children Sample" are boosted by another multiplicative weight adjustment.

This sixth adjustment is given by the inverse of one minus the overall screening rate within a stratum. Note that in P.E.I., this adjustment was implemented a little differently since, among other reasons, the rejective method was applied in all four quarters of data collection rather than in the last two quarters only. Also note that this adjustment was not applied in apartment strata, high income strata and remote strata, since the rejective method was not implemented there.

Note that the non-responding cycle 1 dwellings from the top-up sample were included in this adjustment.

11.1.2.2 Further Adjustments for Selected Members for all Provinces Other Than Québec

The adjusted «stripped» weight for each individual is obtained as follows. First, the «stripped» weight is multiplied by weight adjustments 4 through 6, as well as weight adjustments 7B, 8B and 9B given below.

Data from the selected members' questionnaire are obtained for only one member of a sampled household. If the selected person is a child less than 12 years of age who lives in a "Children" sample dwelling (see Chapter 5, Sample Design for the definition of "Children" sample dwelling) then all of the children in the household to a maximum of four were administered the NLSCY questionnaire in 1994–1995. Otherwise, the selected member was asked an additional set of NPHS questions. Several adjustments have to be made to account for this design and the nonresponse to this questionnaire.

Adjustment 7B: NLSCY Integration Weight Adjustment

In the last two quarters of data collection in 1994–1995, the NPHS selected respondents for both the NPHS and NLSCY selected-member questionnaires. In sampled «Children» households that had children, all children up to a maximum of four (aged less than 12) were selected and administered the NLSCY questionnaire. One child was identified as the NPHS panel member for that household in future cycles. The children's data did not reside on the 1994–1995 NPHS microdata file. For more details on integration with the NLSCY, see Chapter 5, Sample Design. In 1996–1997 and in subsequent cycles, all selected members including the panel children aged less than 12 in 1994–1995 are surveyed by the NPHS, not the NLSCY.

In 1994–1995, households containing children aged 12 or older were selected from the "Adult Sample" only. To compensate for the fact that households containing children coming from the "Children Sample" did not contribute to the estimates for selected individuals in 1994–1995 but will in 1998–1999, the weights for those households containing children sampled in the last two quarters that come from the "Adult Sample" had a special weight adjustment applied. This adjustment is given by the inverse of the proportion of the total sample assigned to the "Adult Sample". For those individuals aged more than 12, one adjustment is made at the cluster level. On the other hand, for those aged 12, a separate adjustment is made for groups of LFS strata (which usually correspond to NPHS strata), to be consistent with Adjustment 9B, which is also made at this level.

Note that the nonresponding cycle 1 dwellings from the top-up sample were excluded from this adjustment.

Adjustment 8B: Selected Member Inverse Selection Probability

As mentioned above, one member from each sampled household is chosen as the selected member. A weight adjustment must be made to reflect the selection and is given by the inverse selection probability. The original intention was that each member would be selected with equal probability given by the inverse of the number of members in the household. However, due to an error in the CAI application, no 12-year-olds were selected in the first two quarters. To compensate, in the last two quarters, instead of each member of a household being selected with the same probability, 12-year-olds were given a larger probability of selection. In P.E.I, 12-year-olds were twice as likely to be selected as any other member aged 13 or more, and elsewhere in Canada, 1.75 times as likely to be selected as any other member aged 13 or more.

Note that the same adjustment was made for the top-up sample of nonresponding cycle 1 dwellings.

Adjustment 9B: 12-year-olds Weight Adjustment

Due to the error mentioned above, 12-year-olds were selected only in the last two quarters of data collection. To obtain an accurate representation of 12-year-olds, their weights had to be adjusted to account for the first two quarters when they had no probability of being selected. This adjustment is made for groups of LFS strata, which usually correspond to NPHS strata, except for the cases of remote and high income strata. In households with children, 12-year-olds could be selected from the "Adult Sample" in all quarters, but were actually only selected from the "Adult Sample" in the last two quarters. Since, within most NPHS strata, 40% of the "Adult Sample" occurred in the last two quarters, the weights of 12-year-olds selected in these two quarters were boosted by the inverse of this rate or by 2.5.

On the other hand, in households with youths but no children, 12-year-olds could be selected from both the "Adult Sample" and the "Children Sample". However, in the first two quarters, they were not selected from the "Adult Sample" as they should have been due to the error mentioned above. Thus, in households with youths but no children, the weights of 12-year-olds were boosted by a multiplicative factor given by the ratio of the percentage of the total sample within an NPHS stratum where they should have been selected to the percentage of the total sample where they were actually selected or by 1.6. Finally, in households with no youths or children, there were no 12-year-olds, so no adjustment was needed. Note that the rates differ somewhat in P.E.I., apartment strata, high income strata and remote strata.

Note also that, due to the same error, it was necessary to include the non-responding cycle 1 dwellings from the top-up sample in this adjustment. Since the selected person for the LFS top-up sample was pre-selected based on the last LFS interview, these households were not affected by this error.

The "weights" referred to above are the «stripped» weights multiplied by all the adjustments to this point (i.e., weight adjustments 4 through 6 as well as 7B, 8B and 9B). These are the 1998–1999 adjusted «stripped» weights for the provinces other than Québec, used to create household member and selected member weights.

11.1.2.3 Adjustments for Québec

Adjustment 4Q: Multiple dwellings Weight Adjustment

Sometimes when an interviewer visited a dwelling, he or she found an extra dwelling that was missed during cluster listing. An example of this might be a basement apartment. In this case each dwelling is known as a multiple. When this occurred, one dwelling was selected at random and interviewed. The weight of the selected dwelling is then adjusted by a multiplicative factor equal to the number of multiples.

Note that this adjustment had to be recalculated to take into account the handful of multiples found in the top-up sample of cycle 1 nonresponding dwellings.

Adjustment 4BQ: Cluster Growth Weight Adjustment

In a few cases, clusters were listed again by NPHS. If there was a growth of 15–30% between ESS counts and NPHS counts, then a multiplicative weight adjustment of

$$\frac{NPHS\ count}{ESS\ count}$$

is made to each selected dwelling within the cluster. If the growth was less than 15%, then the growth is assumed to be negligible and this adjustment is set to one. For all these dwellings, the multiples and cluster growth adjustments are multiplied by the basic dwelling weight to give a "**preliminary weight**".

If the growth was more than 30%, then extra dwellings were selected for NPHS from the extra dwellings listed within the cluster. For these selected extra dwellings, the "**preliminary weight**" is the inverse of the product of the ESS cluster selection probability and NPHS cluster retention probability multiplied by

$$\frac{number\ of\ extra\ dwellings\ listed}{number\ of\ extra\ dwellings\ selected}$$

and the multiples adjustments. Since none of these dwellings was interviewed by the ESS, there is no way to categorize them into one of the ESS household composition categories.

Adjustment 5Q: Household Nonresponse Weight Adjustment

To adjust for total nonresponding households, the following adjustment is made:

$$\frac{\textit{sum of weights for respondong and non - responding households}}{\textit{sum of weights for responding households}}.$$

The weight here is the preliminary weight. A separate adjustment is done within a nonresponse weighting area. For the ESS in-scope dwellings the nonresponse weighting areas are defined as an intersection of an NPHS stratum and ESS household type by quarter. For the dwellings that were added because the cluster had greater than 30% growth during NPHS relisting, the weighting area consists of the added dwellings within the cluster by quarter. The ESS out-of-scope dwellings are grouped into two nonresponse weighting areas by quarter for nonresponse adjustment purposes. The first group contains all dwellings with an ESS response code of 10 (demolished, vacant, abandoned). The second contains all dwellings with an ESS response code of 18 (collective or business). Multiplying the preliminary weight by the household nonresponse weight adjustment produces the "**demographic weight**".

For the top-up sample of nonresponding cycle 1 dwellings, all cycle 3 household nonresponse is dealt with here, in a similar fashion. Note that all out-of-scope dwellings (under construction, rejected households, etc.) are dropped.

11.1.2.4 Further Adjustments for Selected Members for Québec

One member from each responding household is designated as the selected member. If this person is a child less than 12 years of age who lives in a "Children" sample dwelling (see Chapter 5, Sample Design for the definition of "Children" sample dwelling) then all of the children in the household to a maximum of four were administered the NLSCY questionnaire in 1994–1995. Otherwise, the selected member was asked an additional set of NPHS questions. Several adjustments have to be made to account for this design and the nonresponse to this questionnaire.

Adjustment 7BQ: NLSCY Integration Weight Adjustment

In a "Children" sample household where a child is found, one child is chosen to be the selected member for the NPHS longitudinal panel. These children's data did not reside on the 1994–1995 NPHS microdata file. An adjustment has to be made to account for the

adults and youths in these dwellings who had no chance of being the selected member. This adjustment is only applied to adults and youths selected for the longitudinal panel in "Adult" dwellings where children were found by NPHS.

The adjustment is equal to the inverse subsampling rate for the "Adult" sample. The adjustment depends upon whether the ESS found children in the dwelling and upon the ESS urban density class to which the dwelling belongs. A separate adjustment is generated for dwellings where ESS found children and dwellings where ESS did not find children because the subsampling rate was different for these two categories. In the ESS Montreal and regional capitals classes, the adjustment is made at the cluster level, while in the ESS smaller urban agglomerations and rural sector classes, it is made at the NPHS stratum level. For an exception to this rule see "12-year-old Weight Adjustment" later in this section.

Note that the nonresponding cycle 1 dwellings from the top-up sample were excluded from this adjustment.

Adjustment 8BQ: Selected Member Inverse Selection Probability

In a dwelling belonging to the "Children" sample in which there were no children less than the age of 12 or a dwelling belonging to the "Adult" sample, every member was originally intended to have an equal probability of being the selected member. However, due to an error in the data collection computer application, 12-year-olds were not eligible to be selected in the first two quarters. To compensate for this they were given double the probability of being selected in quarters 3 and 4. A weight adjustment equal to the inverse probability of an individual within the household being the selected member is applied.

Note that the same adjustment was made for the top-up sample of nonresponding cycle 1 dwellings. For the LFS top-up sample, each "initially absent" member of the household (i.e., new infants and immigrants) was given an equal probability of selection, based on the household roster at the last LFS interview.

Adjustment 9BQ: 12-year-olds Weight Adjustment

In order to get an accurate representation of 12-year-olds, their weight had to be increased to account for households where they were not eligible to be selected as a result of the software error. This adjustment is equal to the inverse probability that a 12-year-old was eligible to be selected from a dwelling where a person 12 or older

was intended to be the selected respondent. Recall that in the Montreal and regional capital classes, clusters are only covered in one quarter. In quarters 1 and 2, 12-year-olds were not eligible to be selected. Therefore, in order for the weight adjustment to account for these ineligible 12-year-olds, it must be done at the NPHS stratum level rather than the cluster level. For consistency, both the integration and 12-year-old weight adjustment are calculated at the NPHS stratum level for 12-year-olds whatever the ESS class.

Note also that, due to the same error, it was necessary to include the nonresponding cycle 1 dwellings from the top-up sample in this adjustment. Since the selected person for the LFS top-up sample was preselected based on the last LFS interview, these households were not affected by this error.

The "weights" referred to above are the «stripped» weights multiplied by all the adjustments to this point (i.e., weight adjustments 4Q, 4BQ, 5Q, 7BQ, 8BQ and 9BQ). These are the 1998–1999 adjusted «stripped» weights for Québec, used to create household member and selected member weights.

11.1.3 Weight Adjustments for Household Members

Adjustment 10A: Household Nonresponse Adjustment

The definition of a nonresponding household encompasses any of the following situations: refusal, special circumstance, language barrier, no one at home, temporarily absent or computer problem. There are also cases where it was determined that the selected member being followed up in 1998–1999 was dead, institutionalized, had moved to the Yukon or Northwest Territories or out of the country. For cross-sectional weighting purposes, these households are included as responding households at this stage but are subsequently dropped from further calculations. These units do not appear on the cross-sectional microdata files but do appear on the longitudinal file.

To adjust for cases of entire households that did not respond to the 1998–1999 survey, the following adjustment is made:

$$\frac{\textit{sum of weights for responding and non-responding households within weighting class}}{\textit{sum of weights for responding households within weighting class}}.$$

Weighting classes consist of groupings of units (or households) that share the same propensity to respond to the survey. Characteristics from cycle 1, available for cycle 3 respondents and nonrespondents alike, are used to define membership in the weighting classes. Classes are formed using a clustering algorithm that arranges the sample units into a tree structure by successively

splitting the data set into «branches» based on the units' characteristics. Each split aims to divide the units present into two or more groups that are most dissimilar with respect to their observed nonresponse rate (and within which the nonresponse rates are expected to be more similar). A different characteristic may be used to define each split. For example, units may first be divided into owner-occupied dwellings and rented dwellings. The former split may then be further split into five groups based on the level of household income while the latter may be further split based on the respondent's age. Each of the newly formed groups may further be split, based on other characteristics, and so on. The results of the final splits are the weighting classes.

The software *Knowledge Seeker IV for Windows*, developed by ANGOSS Software International Limited, is used to generate the tree structure. We used an improved version of the CHAID (Chi-Square Automatic Interaction Detector) algorithm available in *Knowledge Seeker* to identify at each node the characteristic that best splits the sample into groups that are dissimilar with respect to a certain characteristic, here the response/nonresponse indicator.

When categorical data with more than two levels are used, Knowledge Seeker may group together one or more levels so that the number of «branches» may be less than the number of categories. For continuous characteristics, such as age, *Knowledge Seeker* first divides the data into ten ranges, which may or may not be collapsed, sometimes resulting in only two «branches». Statistical tests are performed at each step to ensure that only statistically significant splits are generated. *Bonferroni* adjustments are made to the significance level of the individual tests to ensure that the significance level of each split is attained. The splitting ends when no more statistically significant splits are found or when splits generate classes that are too small (a minimum of 30 units per class is used). For more information about the CHAID algorithm, see Kass (1980)[5].

Personal characteristics of the selected respondent, as well as dwelling or household characteristics, are used to define the weighting classes for household nonresponse. The selected respondents' personal characteristics are deemed to play a significant role in predicting household nonresponse for several reasons. Often, person and household level nonresponse are equivalent. An obvious example is the fact that selected respondents who could not be located in 1998–1999 led to a household nonresponse situation. Also, during data collection, emphasis was placed on getting a response for the selected respondent, since information on the selected respondent was essential for longitudinal purposes. If the selected respondent was not available or did not respond, the interviewers were instructed not to complete the general component for the rest of the members of the household.

[5] Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29, 119–127.

Finally, in many households, members shared common characteristics, such as race. In those cases, the respondent characteristics are in some sense also household characteristics.

Separate sets of weighting adjustment cells are created for each province. In addition to household and personal characteristics of the selected member, some characteristics that are related to the design of the survey were used, to reduce the effect of the sample design on the results of the statistical inference and incorporate design variables into the analysis. The characteristics vary by province but include the following variables:

Geographic	Province, Census Metropolitan Area, urban/rural indicator
Household	Dwelling type, owner/renter status, family type, household income adequacy, main source of income, nonresponse flag for income in 1994–1995, presence of children in the household
Personal	Sex, age, age over 16 indicator, marital status, race, country of birth, age at immigration, restriction of activity flag, main activity/labour force status

Note that household nonresponse for the nonresponding cycle 1 dwellings from the top-up sample was dealt with in Adjustment 5. LFS top-up sample households were adjusted separately, using only basic classes, due to small sample sizes.

Adjustment 11A: Weight Share Method

Only selected (longitudinal) members are traced between cycles. Therefore if the composition of the household of the longitudinal member changes between cycles, their cohabitants at cycle 2 will not have any weights associated with them since they were not selected in the panel in cycle 1. The weight share method is a mechanism for assigning these individuals weights in such a way that resulting estimates are unbiased.

In cycle 2, every household member was assigned the panel member's weight divided by the number of household members who were in scope to the survey in cycle 1 (e.g., excluding persons born or entering Canada since 1994). For cycle 3, however, because these initially absent persons were accounted for in the LFS top-up sample of immigrants and infants, there was no need to "give" additional weight to these new members. For this reason, a very simple version of the weight share method was used, where the weight is shared equally among all members of the household. In other words, every household member is assigned the panel member's weight divided by the *total* number of household members, whether initially absent in cycle 1 or not.

Note also that, due to an error in the data collection computer application, no 0-11-year-olds were chosen as the selected person for the top-up sample of non-responding cycle 1 dwellings. For this population, the selected member's weight must be *given* to all 0-11-year-olds, and *shared* with other members. In other words, every household member in these households is assigned the selected person's weight divided by the number of household members older than 11. For more information see Ernst (1989)[6] or Lavallée (1995)[7].

Adjustment 12A: Interprovincial Migrations

It is sometimes necessary to make an adjustment for panel members that move from provinces with large populations to those with small populations, as the members' weights are often atypically large compared with those of other similar units in their new province. Corrections are only made in instances where at least one extreme weight was generated within the provincial move pattern in question. Note that corrections were also made for some non movers to reduce extreme weights.

Adjustment 13A: Household Member Nonresponse

Nonresponse at cycle 2 attributable to a household member is less than 2 percent. Therefore, all age–sex groups are combined and the following adjustment is made, by province:

$$\frac{\textit{sum of weights of respondents and non-respondents in a province}}{\textit{sum of weights of respondents in a province}}.$$

Note that top-up households were adjusted separately, using very basic classes.

Adjustment 14A: Integration of Core and Top-up Samples

As mentioned previously, in cycle 3 there is a supplemental top-up sample. However, unlike in cycle 2 with its buy-in samples, there is no need for an explicit integration step with the core sample. The top-up sample of nonresponding cycle 1 dwellings is fully integrated with the core since basic "stripped" cycle 1 weights are available for these units. The other part of the top-up sample, of infant and immigrant households, can be thought of as sampling from subpopulations or conceptual strata that are different from the core population. The core population and these top-up "strata" are mutually exclusive and exhaustive with respect to the Canadian population and can

---

6 Ernst, L. (1989). Weighting Issues for Household and Family Estimates. In *Panel Surveys*. (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh.) New York: John Wiley and Sons, 135–159.

7 Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method. *Survey Methodology*, 21, 25–32

therefore be simply added together for estimation purposes. The only step remaining is to calibrate the survey weights to Canadian population totals.

Adjustment 15A: Poststratification Adjustment

This adjustment ensures that the final weights sum to the 1998–1999 population totals at the province/age/sex level. These totals are based on 1996 postcensal projections. The age-sex categories are people aged 0–11, 12–24, 25–44, 45–64, and 65 and older, males and females. This adjustment is given by

$$\frac{\textit{population projection in a province/age/sex category}}{\textit{sum of weights of respondents in a province/age/sex category}}.$$

Note that households entirely composed of new immigrants since 1994 are now covered by the cross-sectional sample, due to the top-up sample of immigrant households, and are included in the population projections.

Adjustment 16A: Noise Factor

For confidentiality reasons, a «noise» factor has been added to the weights of persons within the same household. This factor follows a uniform distribution and is added in such a way that the sum of the weights at the household level is respected.

The final household member weight on the general file is calculated as the «stripped» weight multiplied by all of the weight adjustments mentioned above. The resulting weight is called the **«Standard Household Member Final Weight» (WT58)**.

11.1.4 Weight Adjustments for Selected Members

The final selected member weight on the health file is calculated as the «stripped» weight multiplied by the following weight adjustments. (Please note that these weight adjustments follow the same numbering scheme as in the previous section, for ease of comparison.)

Adjustment 10B: Household and Selected Member Nonresponse Adjustment

This adjustment compensates for complete household nonresponse, and for selected individuals within responding households who do not answer the selected members' questionnaire. Similar to Adjustment 1A, weighting classes are constructed using information available on the selected member from 1994–1995. Again, separate sets of weighting adjustment cells are created for each province but the input used for constructing the classes is slightly different. These classes are based upon data from all longitudinal respondents, not simply those who had responded to at least the general component in both years, as is

used for the household member nonresponse adjustment. The adjustment is given by

$$\frac{\textit{sum of weights for responding and non - responding households in weighting class}}{\textit{sum of weights for responding households in weighting class}}.$$

Note that household nonresponse for the nonresponding cycle 1 dwellings from the top-up sample was dealt with in Adjustment 5. Selected individual nonresponse for these households, as well as LFS top-up sample households, were adjusted separately, using basic classes, due to small sample sizes.

Adjustment 12B: Interprovincial Migrations

This adjustment is similar to adjustment 12A. In cases where one extreme weight is generated within the provincial move pattern in question, the weights of all those selected members falling within the move pattern in question are reduced so that the sum of their weights equals the demographic projection of the number of movers within the move pattern in the past two years. If the adjustment would have resulted in an increase, the adjustment is not made, since that would augment the extreme weight even further.

Adjustment 14B: Integration of Core and Top-up Samples

As in adjustment 14A, this adjustment is not explicitly needed in 1998–1999.

Adjustment 15B1: Poststratification of 0–11-year-olds

Although the 0–3-year-old population was well represented in the LFS top-up sample of infant households, because of the software error noted in Adjustment 11A, the 4–11-year-old population was not as well represented in the sample of selected persons. This population was represented in the core sample, but not in the top-up sample. To compensate for this lower coverage, a "prepoststratification" step was done before Adjustment 15B2. These two age groupings, 0–3 and 4–11, were separately benchmarked to the equivalent 1998-1999 population totals at the Canada/age/sex level. This adjustment is given by:

$$\frac{\textit{population projection in an age/sex category}}{\textit{sum of weights of respondents in an age/sex category}}.$$

Adjustment 15B2: Poststratification

This adjustment ensures that the final weights on the health file sum to the 1998–1999 population totals at the province/age/sex level. The age-sex categories are people aged 0–11, 12–24, 25–44, 45–64, and 65 and older males and females. This adjustment is given by

$$\frac{\textit{population projection in a province/age/sex category}}{\textit{sum of weights of respondents in a province/age/sex category}}.$$

Note that due to the top-up sample of households with newborns, the first age group once again covers the entire 0–11-year-old age group, unlike in 1996–1997, where 0- and 1-year-olds were excluded. Again, it should be noted that households entirely composed of new immigrants since 1994–1995 are now covered by the 1998–1999 NPHS top-up sample, and are included in the population projections. The resulting weight is called the **«Standard Selected Member Final Weight» (WT68)**.

11.2 Cross-sectional Weighting for the NPHS Cycle 3, Supplemental Top-up Sample

To improve coverage of the 1998-1999 population, and to boost the sample size to counter the effects of attrition, a supplemental sample was selected. This two-part sample is meant only for use in cross-sectional estimation and will only be used for cycle 3. (See Section 5--Sample Design for further details.) To tap into the full cross-sectional potential of the data it is necessary to combine the core NPHS and supplementary top-up data into one large data set and have weights that reflect this combination of data sources. (See Section 12.1, Use of Weights for more details on when to use the various weights.)

The first part of the top-up sample covers the part of the Canadian population that was "initially absent" in 1994-1995. In particular, immigrants entering the country since January 1, 1995, and infants born after that date were selected. The former subsample was selected using a filter question on four rotation groups of the Labour Force Survey, resulting in a sample of over 400 immigrant households. The latter was selected by subsampling from households with 0- to 3-year-old infants from two of these LFS rotation groups, yielding a sample of about 760 infant households.

The second part of the top-up sample was chosen to make up for attrition in the panel that had occurred in the first four years of its life cycle. It was based on nonresponding dwellings from the original sample in cycle 1. For this part of the top-up sample, over 2,800 dwellings were chosen. These include nonrespondents dwellings, refusals, and those with technical problems. Out-of-scope dwellings such as vacant buildings were excluded.

The particular weight adjustments for the LFS top-up sample are noted in this section. The weight adjustments for the nonresponding cycle 1 dwellings are incorporated in Section 11.1, since their initial weights were derived in cycle 1 with the rest of the core household sample.

11.2.1 LFS Basic Weights

As in 1994-1995 (see Section 11.3.1), the starting point was the LFS basic weight, the product of the cluster weight and the dwelling weight.

11.2.2 LFS Subweights

Because the LFS top-up sample required household composition from the last LFS interview, only responding households could be used. To compensate for this reduction in effective sample size, a request was made for the LFS subweights, which have been inflated slightly to account for nonresponse.

11.2.3 Further Weight Adjustments to the Subweights

Adjustment 1T: Rotation Group Weight Adjustment

The full LFS sample is composed of six "rotation groups". Since the top-up sample of infant households was based on only two rotation groups, the multiplicative weight adjustment of 6/2 is made to compensate. In a similar fashion, the top-up sample of immigrant households is adjusted by a factor of 6/4.

Adjustment 2T: Infant Household Subsampling Weight Adjustment

This part of the top-up sample was based on two LFS rotation groups, which yielded more infant households than needed. As a result, about one-half of these households were selected. In order to stabilise the weights at the selected-person level, households with more than one infant were sampled more heavily than one-infant households.

Adjustment 3T: NLSCY New-born Weight Adjustment

In the second of the two LFS rotation groups mentioned above, the available households with new-borns (less than 12 months old) was divided between the NPHS and NLSCY. A second adjustment near 2 was made for these households to compensate for this subsampling. This was done only for the second rotation group, used in the third quarter of NPHS collection.

Adjustment 8T: Selected Member Inverse Selection Probability

For the LFS top-up sample, each "initially absent" member of the household (i.e., new infants and immigrants) was given an equal probability of selection, based on the household roster at the last LFS interview.

The «selected member weight» for the LFS top-up sample is the LFS basic weight multiplied by all the adjustments to this point (i.e., LFS nonresponse adjustment as well as 1T, 2T, 3T and 4T). This weight is used as input to the creation of household member and selected member weights. (See Sections 11.1.3 and 11.1.4.)

11.3 1994–1995-based Weighting Procedures for Provinces Other Than Québec

To begin, the basic LFS weighting procedure, which comprises the first multiplicative weight adjustments necessary in the formation of the «stripped» weights, is described below. Note that the weights were stripped back farther in cycle 3 than in cycle 2, all the way back to Adjustment 3. This was necessitated by the cross-sectional top-up sample, which included nonresponding dwellings from cycle 1. All adjustments from Adjustment 4 onward had to be revisited to properly reintegrate these dwellings into the weighting structure.

11.3.1 LFS Basic Weights

The LFS uses a stratified multistage design (mainly two-stage, but in some cases, three-stage). In both cases, a sample of clusters is selected in each stratum using one of several probability proportional to size (PPS) sampling schemes. An LFS "cluster weight" is then calculated as the inverse probability of selecting a cluster, in accordance with the sample selection scheme. At the next (last) stage, dwellings are selected within sampled clusters using systematic sampling. A "dwelling weight" is calculated as the inverse probability of selecting a dwelling given that the cluster containing it is selected. An "LFS basic weight" is then given by the product of the cluster weight and the dwelling weight.

11.3.2 Further Weight Adjustments to the Basic Weights

Adjustment 1: Rotation Group Weight Adjustment

The full LFS sample is composed of six "rotation groups". The NPHS requested sample from the LFS in terms of integral numbers of rotation groups (between 1 and 6), although a fractional number may actually have been required to fulfil sample size needs (see Adjustment 3 below). Thus, the first multiplicative weight adjustment, which compensates for the integral number requested, is given by:

$$\frac{\textit{number of rotation groups in an LFS stratum used by LFS (usually 6)}}{\textit{intergral number of rotation groups in an LFS stratum requested by NPHS}}.$$

Adjustment 2: Cluster Growth Weight Adjustment

There may be clusters that experience growth between the time when a Census enumeration of the cluster takes place and the time when the cluster is listed for the LFS. The cluster selection probability is based on the Census enumeration figure, which may be out of date. This has the effect that the number of dwellings in the LFS sample increases very slightly with moderate growth in the housing stock. In clusters where substantial growth has taken place, subsampling is used to keep interviewer assignments manageable. The NPHS

instituted a similar subsampling of clusters that had experienced moderate growth. Thus, the second multiplicative weight adjustment is given by the inverse of this subsampling ratio in clusters where subsampling occurred for either the LFS or the NPHS.

Adjustment 3: Stabilization Weight Adjustment

Stabilization is a means of capping the sample size within a stabilization area to prevent the associated costs from becoming too prohibitive. A "stabilization area" consists of clusters in the high-income and apartment frame and of groups of strata in the regular frame. "Stabilization" addresses the problem of growth that occurs within a stabilization area, when the growth is large enough to be a concern even after cluster growth adjustment, although no single cluster has contributed substantially enough to the growth to be considered the root of the problem. This problem is remedied through subsampling within the stabilization area. In addition to regular stabilization, it is at this point that the fractional part of a rotation requested of the LFS, but not required by the NPHS, is also "stabilized out" through subsampling (see Rotation Group Weight Adjustment). Thus, the third multiplicative weight adjustment is given by the following:

$$\frac{\textit{number of dwellings selected by the LFS within a cluster}}{\textit{number of dwellings actually used by the NPHS within a cluster}}.$$

The "weights" referred to above are the LFS basic weights multiplied by all the adjustments to this point (i.e., weight adjustments 1 through 3). These are the 1998–1999 «stripped» weights for the provinces other than Québec, and the starting point for the 1998–1999 weighting.

11.4 1994–1995-based Weighting Procedures for Québec

The National Population Health Survey used a subsample of the *Enquête sociale et de santé* (ESS) in its design (see Chapter 5, Sample Design for more details). For this reason, the calculation of NPHS weights is tied to the weighting procedures used for the ESS. The following sections describe the ESS weighting procedures and the steps required to produce the 1998–1999 «stripped» weights for NPHS members. Note again that the weights were stripped back farther in cycle 3 than in cycle 2, all the way back to the basic dwelling weight. This was necessitated by the cross-sectional top-up sample, which included nonresponding dwellings from cycle 1. All adjustments from this point onward had to be revisited to properly reintegrate these dwellings into the weighting structure.

11.4.1 ESS Weights

The ESS contribution to the weights is calculated as follows:

ESS Cluster Weights

The ESS used a stratified multistage design. After several levels of stratification, clusters were selected from each stratum using probability proportional to size (PPS). The size measure used was the household count in the cluster based upon the 1986 Census. An "**ESS cluster weight**" can be calculated as the inverse probability of selecting a cluster.

ESS Dwelling Weights

After selecting a cluster, a fixed number of dwellings was allocated to be selected from the cluster. Each dwelling in the cluster had an equal chance of being selected. The "**ESS dwelling weight**" is then the inverse of the probability of selecting the dwelling within the cluster multiplied by the ESS cluster weight.

11.4.2 NPHS Basic Dwelling Weights

There were two major steps to selecting the NPHS sample. First, the subset of ESS clusters to be used in the NPHS had to be identified. Second, the subset of ESS dwellings within each retained cluster had to be selected.

Probability of Retaining an ESS Cluster for NPHS

As ESS strata were sometimes very small, NPHS strata were defined as comprising one or more ESS strata. A fixed number of clusters was allocated to be retained from each NPHS stratum. In cases where the NPHS stratum consisted of more than one ESS stratum, the allocation of clusters to ESS strata was proportional to the number of households in each ESS stratum, in order to produce a PPS sample of clusters in each NPHS stratum. Fractional sample sizes were randomly rounded up or down to the next integer. Once the number of clusters to be retained from an ESS stratum had been determined, each cluster within the ESS stratum had the same probability of retention, in most cases. The exceptions were clusters in which the number of dwellings grew by more than 150% between the 1986 Census and the 1992–1993 ESS cluster listing. These clusters were given a higher probability of retention (either 100% or 40% greater probability of retention).

Probability of Retaining an ESS Dwelling for NPHS

In clusters retained for the NPHS, only dwellings selected for the ESS were eligible to be selected for NPHS. Note that those dwellings that were out of scope for the ESS (businesses, collectives, demolished or abandoned) had a

probability of one of being retained, in case they became in scope for NPHS. From the ESS in-scope dwellings, a fixed number of dwellings within each cluster was initially retained for the NPHS. A further subgroup of these selected dwellings was dropped because of their ESS household composition. The probabilities that a dwelling would be retained due to its household composition ranged from one-third for one-person households to 1 for households with children less than 12 years old.

The "**basic dwelling weight**" is the ESS dwelling weight times the inverse of the product of the ESS cluster retention probability and the ESS dwelling retention probability. The ESS dwelling retention probability includes both the probability of a dwelling being initially retained for NPHS and the probability of being retained due to its household composition.

The 1998–1999 "stripped" weight is this basic dwelling weight, and the starting point for the 1998–1999 weighting.

12. File Usage

This section starts with a discussion of the weight variables and explains how they should be used when doing tabulations on the public use microdata files. This is followed by an explanation of the variable naming convention that is employed for all cycles of the NPHS. The last part of the section discusses alternate approaches to data access available to analysts.

12.1 Use of Weights

12.1.1 Cross-sectional Weight, General File WT58

Only one weight, WT58, appears on the general file. This weight is applicable to all age groups and provinces. This weight is based on the total sample, i.e. on the integrated core and supplemental samples. ALL QUESTIONS ON THE GENERAL FILE SHOULD BE ANALYZED USING THIS WEIGHT.

(For a more detailed explanation on the creation of this weight, see sections 11.1 and 11.2 of the documentation on weighting.)

12.1.2 Cross-sectional Weight, Health File WT68

Only one weight, WT68, appears on the health file. This weight is applicable to all age groups and provinces. This weight is based on the total sample, i.e. on the integrated core and supplemental samples. ALL QUESTIONS ON THE HEALTH FILE SHOULD BE ANALYZED USING THIS WEIGHT

(For a more detailed explanation on the creation of this weight, see sections 11.1 and 11.2 of the documentation on weighting.)

12.2 Variable Naming Convention

In 1996-1997, the NPHS adopted a variable naming convention that allows data users to easily use and refer to similar data from different collection periods and across survey components of the NPHS program. The following requirements were mandatory: restrict variable names to a maximum of 8 characters for ease of use by analytical software products; identify the survey occasion (1994-1995, 1996-1997, 1998-1999...) in the name; and allow conceptually identical variables to be easily identifiable over survey occasions. For example, conceptually identical data on smoking were collected in 1994-1995, 1996-1997 and 1998-1999. The variable names for these questions should only differ in the year position identifying the particular survey occasion in which they were collected. This convention will be followed throughout the longitudinal survey, and will be adopted by all NPHS surveys: the household survey, the institutional survey, the Northern survey, and supplements.

12.2.1 Variable Name Component Structure

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

Positions 1-2: Variable / Questionnaire section name
Position 3: Survey type
Position 4: Year/cycle variable appears
Position 5: Variable type
Positions 6-8: Variable number / name from questionnaire

For example: the variables DHC4_AGE, DHC6_AGE, and DHC8_AGE:

DH: in the Demographic and Household content section of the questionnaire;
C: questions which are Core content on the household survey; appeared in 1994-1995 cycle;
4: appeared in 1994-1995 cycle;
6: appeared in 1996-1997 cycle;
8: appeared in 1998-1999 cycle;
_: can be found on the questionnaire;
AGE: the variable name.

12.2.2 Positions 1-2: Variable / Questionnaire Section Name

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

AD	Alcohol dependence	IN	Income
AL	Alcohol	IS	Insurance
AM	Administration (of the survey)	LF	Labour force
AP	Attitudes towards parents	MH	Mental health
BP	Blood pressure	NU	Nutrition
CC	Chronic conditions	PA	Physical activities
CE	Contact exit	PC	Physical check-up
CI	Contact information (institutions, 1998-1999)	PR	Province
DG	Drug use	PY	Psychological resources (self-esteem, mastery, sense of coherence) (Cycles 1 and 3)
DH	Demographics and household	RA	Restriction of activities
DV	Dental visits	RP	Repetitive strain
ED	Education	RS	Road safety
ES	Emergency services	SC	Self-care
EX	Eye examination	SD	Socio-demographics
FH	Family medical history	SH	Sexual health

FI	Food insecurity	SM	Smoking
FL	Balance and falling (institutions)	SP	Sample identifiers (methodology)
FS	Flu shots	SS	Social support
GE	Geographic identifiers (methodology)	ST	Stress
GH	General health	TA	Tobacco alternatives
HC	Health care utilization	TU	Tanning and UV exposure
HH	Household	TW	Two-week disability
HI	Health information	WH	Women's health: breast self-examination, breast examination, mammography and Pap smear
HS	Health status	WT	Weights
HV	HIV		
HW	Height and Weight		
IJ	Injuries		

12.2.3 Position 3: Survey Type

A Asthma supplement
B Province-specific buy-in content - children's questions
C Core questions that will be repeated in each cycle
I Institutions

K Longitudinal children's questions
N North (Yukon / NWT)
P Province-specific buy-in content - adult questions
S National supplement (Health Promotion Survey and Food Insecurity HRDC supplement in 1998)
_ Cycle specific questions, not repeated in every cycle (stress in 1994-1995, access to services in 1996-1997)
3 Survey administration variables for household and demographic component (H03)
5 Survey administration variables for the General component (H05)
6 Survey administration variables for the Health component (H06)

12.2.4 Position 4: Year / Cycle Variable

4 1994-1995
6 1996-1997
8 1998-1999
0 2000-01
2 2002-03
A 2004-05
B 2006-07
C 2008-09
D 2010-11
E 2012-13
F 2014-15

12.2.5 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	Cross-sectional derived variable	A variable calculated from one or more collected or coded variables, usually calculated during head office processing (e.g., health status 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)
L	Longitudinal derived variable	A variable calculated using variables from two or more survey cycles

12.2.6 Positions 6-8: Variable Name

In general, the last three positions follow the naming on the questionnaire. Numbers are used where possible: Q1 becomes 1. «Mark-all» questions use letters for each possible answer category: Q1 (mark all that apply) becomes 1A, 1B, 1C, etc. Demographic variables, which are used frequently by analysts, are identified by a three letter identifier, rather than by a question number; for example «age» is DHC8_AGE in 1998-1999. Where groups of questions with the same topic were collected in sections that had different section names on the questionnaire, position 6 is used to identify the subsection. For example, the first question on chronic stress was named ST_4_C1, the first question on childhood and adult stressors (traumas) was named ST_4_T1. Another

example of this occurs in the general health questions for the Health Promotion Survey in 1996-1997. These questions were separated into three sections for inclusion in the questionnaire and the corresponding variable names reflect this, with position 6 indicating the section in which it appears.

12.3 Access to Master Files 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.

Remote access to the survey master file is one way to reap these benefits. 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 his or her own set of analytical computer programs using the test file to confirm that the routines are functioning correctly. At that point, the code for the custom tabulations is then sent via the e-mail to nphs@statcan.ca. The code will be moved into Statistics Canada's internal secured network and processed against the appropriate master file of NPHS data. 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.

A second approach for any client 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 have access to the master file for the preparation of their own custom calculations. As with remote access, the results are screened for confidentiality and reliability concerns before release. Unlike remote access, there is a charge for this service.

Finally, a Research Program allows researcher to submit to Statistics Canada, a research project that uses data from the Master Files. These projects are accepted based on a set of specific rules. When the project is accepted, the researcher become a Statistics Canada deemed employee and can access the Master Files data from designated STC sites. For more information on this program please contact Mario Bédard by telephone at 1-613-951-8933, by fax at 1-613-951-4198 or by e-mail using the following address: nphs@statcan.ca.