

**Population Health Surveys Program**

**National Population Health Survey**

**Cycle 4 (2000 – 2001), Household Component**

**Longitudinal Documentation**

**May 2002**

## NOTE TO USERS

Starting with the collection of Cycle 4 data in 2000-2001, the National Population Health Survey (NPHS) is strictly longitudinal in nature.

To provide greater flexibility to users, one microdata master file is being issued for NPHS Cycle 4. This file includes all 17,276 NPHS panel members, notwithstanding their response patterns from Cycles 1 to 4. Within the master file, four subsets of respondents have been created with corresponding sampling weights and a flag to make their identification easier.

The NPHS Cycle 4 longitudinal documentation provides a wide range of information on the survey: objectives, content development, sample design, collection, processing, data quality, weighting procedures, tabulation guidelines and data access. Chapters 7 and 9 give more details on the various subsets of respondents and their associated weights.

This guide is also intended for users of the share file. The share file includes the Cycle 4 share respondents and their corresponding sampling weight. This group of respondents is one of the master file subsets of respondents. Users of the share file should disregard references specific to other subsets of respondents.

Finally, this document sometimes refers to a specific cycle of NPHS by using the years in which it occurred. For reference, here is the list of NPHS cycles with their corresponding years:

Cycle 1 = 1994-1995  
Cycle 2 = 1996-1997  
Cycle 3 = 1998-1999  
Cycle 4 = 2000-2001

## **Table of Contents**

	Page
1. Introduction .....	6
2. Background .....	7
3. Objectives .....	8
4. Survey Content .....	9
4.1 Content Selection Criteria .....	9
4.2 Cycle 4 (2000-2001) Changes to Existing Content .....	10
4.3 Cycle 3 (1998-1999) Data Feedback and Follow-up Questions .....	11
5. Sample Design .....	12
5.1 Cycle 1 (1994-1995) Sample Design .....	12
5.1.1 Sample Allocation .....	12
5.1.2 The Rejective Method .....	12
5.1.3 Sample Selection .....	13
5.1.4 Sample Design in Quebec .....	14
5.2 Longitudinal Sample .....	15
6. Data Collection .....	17
6.1 Questionnaire Design and Data Collection Method .....	17
6.2 Tests .....	17
6.3 Interviewing .....	17
6.4 Non-response and Tracing .....	18
7. Data Processing .....	20
7.1 Editing .....	20
7.2 Coding .....	20
7.3 Derived and Grouped Variables .....	20
7.4 Estimation and Weighting .....	21
7.5 Subsets of respondents .....	21
7.6 Definition of Full/Complete response .....	22
8. Response and Attrition Rates .....	23
8.1 Cycle 1 (1994-1995) Response Rates .....	23
8.2 Cycle 2 (1996-1997) Response Rates .....	24
8.3 Cycle 3 (1998-1999) Response Rates .....	25
8.4 Cycle 4 (2000-2001) Response Rates .....	26
8.5 Attrition Rates .....	27
8.6 Non-response Rates Due to Failure to Trace .....	28

9.	Weighting.....	29
9.1	Longitudinal Weighting.....	29
9.1.1	Cycle 1 (1994-1995) Stripped Weights.....	30
9.1.2	Longitudinal Weights.....	30
9.1.2.1	Longitudinal Square Weight (WT64LS).....	30
9.1.2.2	Longitudinal Full Weight (WT60LF).....	30
9.1.2.3	Longitudinal Full Weight for Cycles 1 and 4 (WT60LFE).....	35
9.1.2.4	Longitudinal Full Share Weight (WT60SLF).....	36
10.	Data Quality.....	37
10.1	Survey Errors.....	37
10.1.1	Non-sampling Errors.....	37
10.1.2	Sampling Errors.....	38
10.1.2.1	Bootstrap Method for Variance Estimation.....	38
10.1.2.2	Bootvar Program for Variance Estimation.....	39
11.	Guidelines For Tabulation, Analysis and Release.....	40
11.1	Rounding Guidelines.....	40
11.2	Sample Weighting Guidelines for Tabulation.....	41
11.2.1	Definitions of Types of Estimates: Categorical vs. Quantitative.....	41
11.2.2	Tabulation of Categorical Estimates.....	41
11.2.3	Tabulation of Quantitative Estimates.....	42
11.3	Guidelines for Statistical Analysis.....	42
11.4	Release Guidelines.....	43
12.	Using the Longitudinal Master Files.....	44
12.1	Use of Longitudinal Weights.....	44
12.2	Ensuring the Reliability of Estimates with the Use of Bootstrap Weights.....	44
12.3	Variable Naming Convention.....	45
12.3.1	Variable Name Component Structure.....	46
12.3.2	Positions 1-2: Variable name / Questionnaire Section Name.....	46
12.3.3	Position 3: Survey Type.....	48
12.3.4	Position 4: Year / Cycle Variable.....	48
12.3.5	Position 5: Variable Type.....	49
12.3.6	Positions 6-8: Variable Name.....	49
13.	Access to NPHS Data.....	50
13.1	Microdata Files.....	50
13.2	Analytical Reports and Tabulations.....	50

**Tables list**

Table 5.A	Longitudinal Sample Size by Province
Table 7.A	Subsets of Respondents
Table 8.A	Tracing for Cycles 2, 3 and 4
Table 9.A	Subsets of Respondents and Corresponding Sampling Weights and Flags
Table 9.B	Variables for Cycle 2 / Non-response Adjustment - WT60LF
Table 9.C	Variables for Cycle 3 / Non-response Adjustment - WT60LF
Table 9.D	Variables for Cycle 4 / Non-response Adjustment - WT60LF
Table 11.A	Sampling Variability Guidelines
Table 12.A	Subsets of respondents and corresponding Bootstrap weights files
Table 12.B	Longitudinal Variables

**Appendix list**

Appendix A:	NPHS Household Component, Health Content over 4 cycles
Appendix B:	NPHS Household Component, Changes to the questionnaire for Cycle 4 (2000-2001)
Appendix C:	NPHS Household Component, Examples of Cycle 3 (1998-1999) Data Feedback and Follow-up Questions,

**Other Reference Documents**

Questionnaire
Record Layout
Alphabetic Index
Topical Index
Data Dictionary
NPHS Derived Variables Documentation, Cycles 1 to 4

## **1. Introduction**

The National Population Health Survey (NPHS) is designed to collect information on the health of the Canadian population and related socio-demographic information. The first cycle of data collection took place in 1994-1995. The survey will continue every second year thereafter for 20 years. The NPHS fulfilled both cross-sectional and longitudinal needs during its first three cycles, but with Cycle 4 (2000-2001) NPHS became strictly a longitudinal survey. The cross-sectional component of the Population Health Survey Program has been taken over by the Canadian Community Health Survey (CCHS). The NPHS is now composed of two components: the survey of households and the survey of health care institutions. Starting with Cycle 4, the survey of the North will be strictly cross-sectional and will be conducted by CCHS rather than NPHS.

The NPHS - Household component includes household residents in all provinces, with the principal exclusion of populations on Indian Reserves and Crown Lands, residents of health care institutions, full-time members of the Canadian Armed Forces, and some remote areas in Ontario and Quebec. The Health Care Institutions component includes long-term residents (expected to stay longer than six months) in health care facilities with four or more beds in all provinces with the principal exclusion of the territories and full-time members of the Canadian Armed Forces. The Household component of NPHS has completed four release cycles: NPHS Cycle 1 (1994-1995), NPHS Cycle 2 (1996-1997), NPHS Cycle 3 (1998-1999) and NPHS Cycle 4 (2000-2001).

The Cycle 4 NPHS Household component collected in-depth information on the health of the longitudinal respondent who was randomly selected in Cycle 1 and also demographic information about all members of the longitudinal respondent household. The health questionnaire includes questions related to health status, use of health services, determinants of health, chronic conditions and activity restrictions. Socio-demographic information is also collected; it includes age, sex, education, ethnicity, household income and labour force status.

This document has been produced to facilitate the manipulation of the Cycle 4 Longitudinal Master File from the Household component containing the survey results. This file is described in more detail in the following chapters. Any questions about the data sets or their use should be directed to:

Survey content and remote access support:  
NPHS Section, Health Statistics Division

Tel: 1-613-951-1653  
Fax: 1-613-951-4198  
e-mail : [nphs-ensp@statcan.ca](mailto:nphs-ensp@statcan.ca)

Custom tabulations/general data support call:  
Client Custom Services, Health Statistics Division

1-613-951-1746  
e-mail : [hd-ds@atstatcan.ca](mailto:hd-ds@atstatcan.ca)

## **2. Background**

In the Fall of 1991, the National Health Information Council 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 system 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 utilization, 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;
- allow the possibility of linking survey data to routinely collected administrative data such as vital statistics, environmental measures, community variables, and health services utilization.



## **4. Survey Content**

The objectives described in Chapter 3 provided a broad direction for 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 Cycle 4 (2000-2001) as well as new content for Cycle 4. The last section introduces Cycle 3 (1998-1999) variables that were fed back and used in Cycle 4. More detailed information on the Cycle 4 content as well as a summary of changes to the content through the four cycles are available in Appendices A, B and C.

### **4.1 Content 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 system.
- 6) The survey should collect information on factors related to good health, not just those related to illness.

During the first three cycles, in each household some limited information was collected from all household members (General component, H05, e.g., socio-demographic characteristics and labour force activity) and one person in each household was randomly selected as the longitudinal respondent for a more in-depth interview (Health component, H06, e.g., health status). For Cycle 4, the longitudinal respondent was the only person who provided all the answers to General and Health component questionnaires.

Reflecting the above criteria, the questionnaire includes questions related to health status, use of health services, determinants of health, chronic conditions and activity restrictions, 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 (traditional and non-traditional), hospital care and on use of drugs and other medications. Health determinants that were explored included smoking, alcohol use and physical activity. Questions were asked on preventive tests and examinations, which probed for frequency and reasons for use. Demographic and socio-economic information include age, sex, education, ethnicity, race, household income and labour force status.

A copy of the NPHS Cycle 4 (2000-2001) questionnaire is provided with the documentation and is also available on Statistics Canada's Web site [www.statcan.ca](http://www.statcan.ca), choose Products and Services, Free Publications, Health and then select the entry for the National Population Health Survey. Please see Chapter 13 for more details on information available on Statistics Canada's Web site.

#### **4.2 Cycle 4 (2000-2001) Changes to Existing Content**

With the introduction of the Canadian Community Health Survey (CCHS), there were many changes to the 2000-2001 National Population Health Survey - Household questionnaire. Since NPHS is strictly a longitudinal survey, some content was migrated to the CCHS (such as the two-week disability section and certain questions on place where health care was provided) or was dropped (e.g. certain chronic conditions), while the order of the questionnaire changed. Follow-up questions were modified and some constant questions (e.g. country of birth) were dropped. Improvements were made to core content and wording.

As only the longitudinal respondent is now surveyed, it was no longer necessary to distinguish between the General questionnaire (socio-demographic characteristics and certain health questions for all household members) and the Health component (more detailed health questionnaire for the longitudinal respondent) components. Therefore the two sets of questions were combined into one large questionnaire. Demographic information on all household members is still gathered (age, sex, relationships between household members).

The order of the questionnaire was changed using a strategy similar to the ordering of the General component in previous cycles (and similar to other social surveys such as the General Social Survey): health data first and socio-economic data (language, ethnicity, race, education, labour force status, and income) at the end of the survey. There was no buy-in content in this cycle.

As always for each cycle, there are focus questions. The major focus of Cycle 4 was stress. Stress questions from Cycle 1 were brought back in Cycle 4. They covered ongoing problems, recent life events, childhood and adult stressors, work stress, self-esteem and mastery. A new question on self-perceived stress was added to the general health section. In addition, more detailed questions about arthritis, diabetes and heart disease were introduced as focus content.

The Injuries section was modified to include new content and to bring categories in line with ICD -10 (International Statistical Classification of Diseases and Related Health Problems, 10th Revision). Major revisions to the Labour Force section were also made in order to include new content and to bring concepts in line with those of the Labour Force Survey.

For more information, please see Appendices A and B.

#### **4.3 Cycle 3 (1998-1999) Data Feedback and Follow-up Questions**

In order to reduce respondent burden, questions to which the answer was already known and that would not change over time (e.g., country of birth) were not asked again. 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 4 interview. This proved to be a valuable tool resulting in better quality estimates. For instance, previous information on selected chronic conditions was fed back in order to explain any change between Cycle 3 and Cycle 4 interviews. For more information please see Appendix C.

## 5. Sample Design

The target population of the NPHS Household component includes household residents in all Canadian provinces in 1994-1995 excluding persons living on Indian Reserves and Crown Lands, residents of health care institutions, full-time members of the Canadian Armed Forces and some remote areas in Ontario and Quebec. This chapter describes the Cycle 1 sample design and explains how the longitudinal sample (17,276 persons) was obtained.

### 5.1 Cycle 1 (1994-1995) Sample Design

The Labour Force Survey (LFS), redesigned in 1991, was used as the basis for the sample design in all provinces except Quebec where the NPHS selected a sample from households already being interviewed by Santé Québec for the 1992-1993 *Enquête sociale et de santé* (ESS).

Three factors shaped the sample 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 representativity, and the selection of provincial samples outside Quebec.

#### 5.1.1 Sample Allocation

The NPHS initially had a target 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)<sup>1</sup>. According to this scheme the sample was allocated proportionally to  $\sqrt{(0.804W_h^2 + 1/12^2)}$ , where  $W_h$  is the 1991 Census proportion of households in province/territory ( $h, h=1, \dots, 12$ ). This allocation determined the base sample size for each province. Four provinces chose to increase their allotted sample size for the first cycle through the buy-in of additional units, but only for cross-sectional purposes. These additional units were not retained for the longitudinal sample.

#### 5.1.2 The Rejective Method

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

---

<sup>1</sup> Kish, L. (1988). Multipurpose Sample Design, *Survey Methodology*, 14, 19-32.

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 method 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 out of 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 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 percentages 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 method was also not applied in remote regions because of the high contact costs there.

### **5.1.3 Sample Selection**

The sample design considered for the household component of the NPHS was a stratified multi-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, or households, were selected from the lists.

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

The basic LFS design is a multi-stage 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), were selected with Probability Proportional to Size (PPS). In a few cases where the population density was low an additional stage was 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 was 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 after listing operations in sampled clusters were completed. As sampling rates were predetermined there were often

differences between anticipated and obtained sample counts. Excessive sample yields were corrected by dropping a portion of the originally selected units. This was usually done at aggregated levels and was called sample stabilization. Note also that sample sizes were inflated to represent dwellings rather than households, as a certain amount of non-response was expected, and a portion of the dwellings were 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 rotations per province, a rotation being one-sixth of the total sample. Sample stabilization is used to maintain the sample at a desired level, as when two rotations are reserved but the sample size needed only represents 1.5 rotations.

Requirements specific to the NPHS led to two modifications to this sampling strategy. The number of "reserves" needed was specified at the stratum level rather than the provincial level in order to meet the specific sub-provincial sample size requirements for cross-sectional purposes in the first cycle. It was also required that the number of clusters selected per stratum be a multiple of four for variance estimation and seasonal representativity (this allowed strata to have two or more independent samples of four clusters each—one per collection period). As 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. Once strata were grouped, their sample clusters were also grouped to form replicates.

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. Two design variables named "Stratum" and "Replicat" can be found on the Master file, where Stratum represents the LFS stratum, and Replicat represents the NPHS replicates.

#### **5.1.4 Sample Design in Quebec**

In Quebec the NPHS sample was selected from dwellings participating in a Santé Québec health survey: the 1992-1993 *Enquête sociale et de santé* (ESS). The survey sampled 16,010 dwellings using a two-stage sample design similar to that of the LFS. The province was divided geographically by crossing 15 health areas with four urban density classes (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: 10 per cluster in major cities, 20 or 30 elsewhere.

Santé Québec provided non-confidential information which allowed the classification of their 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). A household type was determined by NPHS personnel 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  $\sqrt{(2W_h^2 + 1/4^2)}$ , where  $W_h$  is the population share for class  $h$ ,  $h=1,2,3,4$ . In each class, an attempt was made to obtain a sub-sample from the ESS, which, 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, which was almost 50% higher than needed, was thus selected. After removing from it 2/3 of the one-member households, 1/2 of the other households with no youths or children, and 1/6 of households with youths but no children, the objective was nearly attained.

Considerations for seasonal representation and variance estimation, and integration with the National Longitudinal Survey of Children and Youth, affected the sub-sampling in Quebec 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, as sample sizes are higher there). The sample of households with children was split into an "Adult" sample and a "Children" sample by a 3:2 ratio, the terms having the same meaning as in other provinces. "Children" sample households in quarters 1 and 2 were reassigned to quarters 3 and 4. Since 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 "Children" sample.

## **5.2 Longitudinal Sample**

The longitudinal sample, also called the longitudinal panel or simply the panel, is composed of the 17 276 persons that were selected in Cycle 1 and had completed at least the General component of the questionnaire in Cycle 1. This panel was surveyed in Cycles 2, 3 and 4 and will be surveyed in future NPHS cycles. Additional samples added to Cycles 1, 2 and 3 for cross-sectional purposes are not part of the longitudinal sample.

The longitudinal sample is not renewed over time. No panel members were or are to be classified out-of-scope. The longitudinal sample size remains the same (17,276) for all cycles. Consequently, for Cycle 4, all longitudinal panel members were 6 years old and over and the longitudinal sample did not contain anyone who has immigrated to Canada after 1994-1995.

The number of people answering the survey slightly decreases from one cycle to the next due to attrition caused by non-respondents, refusals and individuals that were untraceable. Despite the attrition, the longitudinal sample is still representative of the 1994-1995 population. The attrition being relatively small (see Section 8.5), it should not lead to large increases in the variance of estimates. Note that panel members who died and panel members who moved to a health care institution are still part of the longitudinal sample and are considered as respondents. Therefore, these persons do not contribute to the attrition of the NPHS longitudinal panel.

Table 5.A presents the sample size of the longitudinal sample by province in 1994-1995. It also shows the number of people that responded to all four cycles of NPHS (this count includes the deceased persons).

**Table 5.A: Longitudinal Sample Size by Province**

<b>Province</b>	<b>Longitudinal Sample Cycle 1 (1994-1995)</b>	<b>Records Providing a Full Response in Cycles 1, 2, 3 and 4</b>
Newfoundland	1,082	905
Prince Edward Island	1,037	856
Nova Scotia	1,085	843
New Brunswick	1,125	893
Quebec	3,000	2,393
Ontario	4,307	3,236
Manitoba	1,205	996
Saskatchewan	1,168	985
Alberta	1,544	1,189
British Columbia	1,723	1,286
<b>Total</b>	<b>17,276</b>	<b>13,582</b>



## **6. Data Collection**

### **6.1 Questionnaire Design and Data Collection Method**

The survey questions were designed for computer-assisted interviewing (CAI), which means that, as the questions were developed, the associated logical flow into and out of the questions was specified, along with 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 is 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 advantage is automatic insertion of reference periods based on current dates. Pre-filling of text or data based on information gathered during the current interview or previous cycles interviews 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 is customised to the respondent based on the data collected.

### **6.2 Tests**

The CAI application was extensively tested in-house in order to identify any errors in the program flow and text. Furthermore, in each cycle, two field tests were conducted. The tests involved four of Statistics Canada's Regional Offices. 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 response rates and to test feedback questions. Field operations and procedures, interviewer training, and the CAI application (i.e., the questionnaire on computer) were also tested. Application testing was an ongoing operation up until the start of collection for the Household survey.

### **6.3 Interviewing**

In Cycle 4, collection for the household sample was divided into four quarters (starting in June, August and November 2000, and February 2001). An additional collection period was held in June 2001 with further tracing attempts of non-respondents from previous quarters.

The interviewers were part-time employees hired and trained specifically to carry out surveys using computer-assisted interviewing, and most were experienced Labour Force Survey (LFS) interviewers. The LFS supervisory and control structure was employed for the NPHS collection. All interviewers attempted a training session that focused on NPHS content and they received an Interviewer's manual for use as a reference tool.

Each living longitudinal panel member received by mail a letter announcing the start of NPHS Cycle 4 data collection. They also received a brochure that presents some results extracted from the survey data.

In general, respondents in the household sample are contacted by telephone, and over 95% of the interviews in Cycle 4 were done over the 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. The total interview time averaged one hour per household.

In all dwellings, information about all household members (age, sex, and relationships between members) was obtained from the person at home at the time when the interviewer called. Proxy reporting for the longitudinal respondent was allowed only for reasons of illness or incapacity. Such proxy reporting accounted for 4.7% of the information collected for respondents aged 12 years and older. On the other hand, almost all interviews for 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 longitudinal respondents. 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. For individuals who 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 co-operation. This was followed by a second call (or visit) from the interviewer.

Many strategies were put in place to reduce the number of non-response cases. The maximum assignment size for an interviewer was set to avoid overburdening interviewers and was based on the experience from previous cycles. This allowed for the efficient follow-up of non-contact cases. 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 to participate in the survey. To maximise 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 another type of non-response. Interviewers used several methods 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 a previous cycle. 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 non-response rate due to failure to trace the longitudinal respondent is 2.9% for Cycle 4, which is relatively low.

Attempts were made to contact panel members who moved within Canada or to the United States. For panel members living outside Canada and the United States, attempts were made to confirm their place of residence. The survey was not conducted if these members were still living outside Canada and the United States.

## **7. Data Processing**

### **7.1 Editing**

Editing was performed on-line in the CAI application during data collection. It was not possible to enter out-of-range values, and flow errors were controlled through the use of CAI. For example, CAI ensured that questions that did not apply to the respondent were not asked. In the case of contradictory responses between questions, warning messages were invoked. In some situations the conflict had to be resolved before the interview could continue. In other situations no corrective action had to be taken. Because of such cases, 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 to either new unique categories, or to a 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.

On the Cycle 3 file, the industry and occupation data from Cycle 3 were coded to the North American Industrial Classification System (NAICS) and Standard Occupational Classification 1991 (SOC-91) while the codes for the Cycle 1 and Cycle 2 industry and occupation data remained coded to the Standard Industrial Classification 1980 (SIC-80) and the Standard Occupational Classification 1980 (SOC-80). In the Cycle 4 master data file (which contains all longitudinal panel members), the Cycle 1 and Cycle 2 industry and occupation data were re-coded, and thus the data for all four cycles is coded to NAICS and SOC-91.

### **7.3 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 variable names generally have a "D" in the fifth character of the variable name (see Section 12.3 for more detail on the variable naming conventions). 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. See the document called "National Population Health Survey – Derived Variables Documentation – Cycles 1 to 4" for the details on how these variables were derived.

Grouped variables were created from certain variables; i.e. the values of the variable have been grouped in order to create another variable. Grouped variable names generally have a "G" in the fifth character of the variable name (see Section 12.3 for more detail on the variable naming conventions).

#### **7.4 Estimation and 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 which calculates, for each person, his or her associated weight. This weight must be used to derive meaningful estimates from the survey. For example, if the number of individuals whose general health has deteriorated between the two cycles of the survey 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.

The NPHS weighting method is presented in Chapter 9.

#### **7.5 Subsets of respondents**

In order to provide greater flexibility to users, a single microdata master file is being issued for NPHS Cycle 4. This file includes all 17,276 NPHS panel members, notwithstanding their response patterns from Cycles 1 to 4. Within the master file, four subsets of respondents have been created along with corresponding sampling weights and a flag to make their identification easier. Refer to Chapter 9 for more information regarding the calculation of each subset's sampling weights and to Section 12.1 for the use of longitudinal weights. Table 7.A provides a description of the four subsets of respondents based on the type of response.

**Table 7.A: Subsets of Respondents**

<b>Subset of Respondents</b>	<b>Type of Response</b>	<b>Number of Respondents</b>
Longitudinal Square	Complete panel: all panel members regardless of their response pattern.	17,276
Longitudinal Full	All panel members with a complete response (Full) in Cycles 1, 2, 3 and 4.	13,582
Longitudinal Full C1 and C4	All panel members with a complete response (Full) in Cycles 1 and 4 (C1 and C4) regardless of their response pattern in Cycles 2 and 3.	14,321
Longitudinal Full Share	All panel members with a complete response (Full) in Cycles 1, 2, 3 and 4 and who agreed to share their data in Cycle 4.	13,110

Users of the share file should note that the “Longitudinal Full Share” subset of respondents is provided separately on a CD-ROM with the corresponding sampling weights. The sampling weights and the flags of the other subsets are not on the share file CD-ROM.

## **7.6 Definition of Full/Complete response**

The last three subsets of respondents have been created using the definition of what is called a “Complete/Full response”. Since Cycle 4, NPHS is strictly longitudinal. The definition of a response is not the same for longitudinal and cross-sectional purposes. For the NPHS longitudinal panel; a Full/Complete response includes:

- **Panel members** who provided a complete response to the interview (i.e., answered all the questions up to a given point in the NPHS questionnaire).
- **Deceased panel members.** The death of a longitudinal panel member is confirmed against the Canadian Vital Statistics Database – Deaths when possible. When the death is confirmed, the cause of death is captured. Variables for panel members who died are set to “9” (i.e., not stated) in the dataset.
- **Institutionalized panel members.** They were surveyed through the NPHS Health Care Institution survey. Their data were brought back to the NPHS household component dataset. However, the institution component collects less information than the household component. The missing variables are set to “6” (i.e., not applicable).

## **8. Response and Attrition Rates**

This chapter presents the response rates and describes how they are computed. The calculation of Cycle 1 response rates is not the same as the calculation of the response rates for the other cycles. Cycle 1 response rates are based on the 20,095 in-scope persons selected to form the panel while response rates for subsequent cycles are based on the 17,276 individuals who form the longitudinal panel members. Another important difference: for the first three cycles, the selected-person response rate is calculated both for the General component (H05) and for the Health component (H06). Since the survey became purely longitudinal in Cycle 4 and there was no longer a distinction between these two components, the longitudinal panel response rate has been calculated only once for Cycle 4.

### **8.1 Cycle 1 (1994-1995) Response Rates**

Cycle 1 response rates are based on the 20,095 in-scope persons selected to form the panel. Consequently, persons who were part of an out-of-scope household (status code = 017, 018, 023, 024)<sup>2</sup> were excluded from the panel and from the calculations of the Cycle 1 response rates.

#### ***Selected-person response rate for H05***

$$\frac{\text{\# of selected persons responding to the H05 component}}{\text{all in-scope selected persons}}$$

The selected-person response rate for the H05 component at the Canada level for the NPHS was **86.0%**. At the provincial level, this rate varied from 80.7% in Ontario to 91.0% in Alberta.

#### ***Selected-person response rate for H06***

$$\frac{\text{\# of selected persons responding to the H06 component}}{\text{all in-scope selected persons}}$$

The selected-person response rate for the H06 component was **83.6%** at the Canada level, and ranged from 77.8% in Ontario to 89.1% in Alberta.

---

<sup>2</sup> 017 = Other ineligible dwelling (e.g., embassy).  
018 = Rejected household.  
023 = Under construction or demolished.  
024 = Dwelling vacant.

***Relevant information for calculation of response rates:***

Number of in-scope selected persons:	20,095
Number of respondents for H05 at the selected-person level:	17,276
Number of respondents for H06 at the selected-person level:	16,794
Number of non-respondents for H05 at the selected-person level:	2,819
Number of non-respondents for H06 at the selected-person level:	3,301

***Calculation of the selected-person response rates:***

$$\begin{aligned} \text{Selected-person H05 Rate} &= \frac{17,276}{17,276 + 2,819} = \frac{17,276}{20,095} = 86.0\% \\ \text{Selected-person H06 Rate} &= \frac{16,794}{16,794 + 3,301} = \frac{16,794}{20,095} = 83.6\% \end{aligned}$$

**8.2 Cycle 2 (1996-1997) Response Rates**

All Cycle 2 response rates are based on the 17,276 individuals who form the longitudinal panel. Persons who have died or been institutionalized are counted as a response for longitudinal purposes; no panel members are classified as out-of-scope.

***Panel response rate for H05***

$$\frac{\text{\# of panel members responding to the H05 component or who have died or been institutionalized}}{\text{\# of panel members}}$$

At the Canada level, the panel response rate for the H05 component was **93.6%**. At the provincial level, this rate varied from 90.4% in British Columbia to 96.2% in Newfoundland.

***Panel response rate for H06***

$$\frac{\text{\# of panel members responding to the H06 component or who have died or been institutionalized}}{\text{\# of panel members}}$$

The panel response rate for the H06 component was **92.8%** at the Canada level. At the provincial level, this rate varied from 89.6% in British Columbia to 95.1% in Newfoundland.



***Relevant information for calculation of response rates:***

Number of longitudinal panel members:	17,276
Number of panel members who have died <sup>3</sup> :	287
Number of panel members who have been institutionalized:	62
Number of respondent panel members for H05:	15,819
Number of respondent panel members for H06:	15,687
Number of non-respondent panel members for H05:	1,108
Number of non-respondent panel members for H06:	1,240

***Calculation of the panel response rates:***

$$\text{H05 response rate for panel} = \frac{15,819 + 287 + 62}{15,819 + 287 + 62 + 1,108} = \frac{16,168}{17,276} = 93.6\%$$

$$\text{H06 response rate for panel} = \frac{15,687 + 287 + 62}{15,687 + 287 + 62 + 1,240} = \frac{16,036}{17,276} = 92.8\%$$

**8.3 Cycle 3 (1998-1999) Response Rates**

As for Cycle 2, the Cycle 3 longitudinal response rates are based on the entire longitudinal panel. Persons who have died or been institutionalized are counted as a response for longitudinal purposes; no panel members are classified as out-of-scope.

***Panel response rate for H05***

$$\frac{\text{\# of panel members responding to the H05 component or who have died or been institutionalized}}{\text{\# of panel members}}$$

At the Canada level, the panel response rate for the H05 component was **88.9%**. At the provincial level, this rate varied from 84.2% in British Columbia to 92.5% in Newfoundland.

***Panel response rate for H06***

$$\frac{\text{\# of panel members responding to the H06 component or who have died or been institutionalized}}{\text{\# of panel members}}$$

The panel response rate for the H06 component was **88.2%** at the Canada level. At the provincial level, this rate varied from 83.9% in British Columbia to 92.0% in Newfoundland.

---

<sup>3</sup> At that moment, the deaths were not confirmed with the Canadian Vital Statistics Database – Deaths, and four of them were found to be alive in Cycle 3.

***Relevant information for calculation of response rates:***

Number of longitudinal panel members:	17,276
Number of panel members who have died <sup>4</sup> :	599
Number of panel members who have been institutionalized:	114
Number of respondent panel members for H05:	14,647
Number of respondent panel members for H06:	14,532
Number of non-respondent panel members for H05:	1,916
Number of non-respondent panel members for H06:	2,031

***Calculation of the panel response rates:***

$$\begin{aligned} \text{H05 response rate for panel} &= \frac{14,647 + 599 + 114}{14,647 + 599 + 114 + 1,916} = \frac{15,360}{17,276} = 88.9\% \\ \text{H06 response rate for panel} &= \frac{14,532 + 599 + 114}{14,532 + 599 + 114 + 2,031} = \frac{15,245}{17,276} = 88.2\% \end{aligned}$$

**8.4 Cycle 4 (2000-2001) Response Rates**

As for Cycles 2 and 3, the Cycle 4 longitudinal response rate is based on the 17,276 members of the longitudinal panel. Persons who have died or been institutionalized are counted as a response for longitudinal purposes; no panel members are classified as out-of-scope. As of Cycle 4, NPHS is now purely longitudinal and no longer distinguishes the H05 questionnaire from the H06 questionnaire; only one response rate is calculated.

***Response rate***

$$\frac{\text{\# of panel members responding or who have died or been institutionalized}}{\text{\# of longitudinal panel members}}$$

The panel member response rate was **84.8%** at the Canada level. At the provincial level, this rate varied from 80.5% in British Columbia to 90.9% in Saskatchewan.

---

<sup>4</sup> At that moment, the deaths were not confirmed with the Canadian Vital Statistics Database – Deaths, and one of them was found to be alive in Cycle 4.

***Relevant information for calculation of response rates:***

Number of longitudinal panel members:	17,276
Number of panel members who have died:	957
Number of panel members who have been institutionalized:	135
Number of respondent panel members:	13,559
Number of non-respondent panel members:	2,625

***Calculation of the panel response rate:***

$$\text{Response rate for panel} = \frac{13,559 + 957 + 135}{13,559 + 957 + 135 + 2,625} = \frac{14,651}{17,276} = 84.8\%$$

**8.5 Attrition Rates**

Attrition is a loss in sample size due to non-respondents, movements out-of-scope and untraceable individuals. Two different attrition rates can be calculated: one showing the attrition rate observed at the end of each cycle, the other showing the cumulative attrition rate based on the original sample. Both of these rates are calculated using the number of individuals found in the Full subset of respondents (for more see Section 7.5).

***Relevant information for calculation of attrition rates:***

Number of longitudinal panel members:	17,276
Number of individuals in the Cycle 2 (1996-1997) Full subset	15,670
Number of individuals in the Cycle 3 (1998-1999) Full subset	14,619
Number of individuals in the Cycle 4 (2000-2001) Full subset	13,582

***Attrition rates at the end of each cycle***

Cycle 2 (1996-1997):	$\frac{17,276-15,670}{17,276}$	=	$\frac{1606}{17,276}$	=	9.3%
Cycle 3 (1998-1999):	$\frac{15,670-14,619}{15,670}$	=	$\frac{1051}{15,670}$	=	6.7%
Cycle 4 (2000-2001):	$\frac{14,619-13,582}{14,619}$	=	$\frac{1037}{14,619}$	=	7.1%

***Cumulative attrition rates***

Cycle 2 (1996-1997):	$\frac{17,276-15,670}{17,276}$	=	$\frac{1606}{17,276}$	=	9.3%
Cycle 3 (1998-1999):	$\frac{17,276-14,619}{17,276}$	=	$\frac{2657}{17,276}$	=	15.4%
Cycle 4 (2000-2001):	$\frac{17,276-13,582}{17,276}$	=	$\frac{3694}{17,276}$	=	21.4%

**8.6 Non-response Rates Due to Failure to Trace**

The failure to trace a longitudinal panel member is a type of non-response. Despite the numerous efforts from the interviewers (discussed in Section 6.4), the cumulative unable-to-trace rate is increasing with the passing cycles but many attempts were put in place to keep this rate as low as possible. Table 8.A contains a summary of the tracing for Cycles 2, 3 and 4.

**Table 8.A: Tracing for Cycles 2, 3 and 4**

		Cycle 4			Total
		Traced and respondent	Traced and non-resp.	Not traced	
Cycle 2 Traced and respondent		14,426	1,413	320	16,159
	Cycle 3 Traced and respondent	14,071	900	189	15,160
	Cycle 3 Traced and non-resp.	303	504	52	859
	Cycle 3 Not traced	52	9	79	140
Cycle 2 Traced and non-respondent		202	587	32	821
	Cycle 3 Traced and respondent	151	29	6	186
	Cycle 3 Traced and non-resp.	44	553	9	606
	Cycle 3 Not traced	7	5	17	29
Cycle 2 Unable to trace		103	41	152	296
	Cycle 3 Traced and respondent	60	10	11	81
	Cycle 3 Traced and non-resp.	4	10	8	22
	Cycle 3 Not traced	39	21	133	193
<b>Total</b>		<b>14,731</b>	<b>2,041</b>	<b>504</b>	<b>17,276</b>

For example, among the 15,160 panel members who were traced respondents in both Cycle 2 and Cycle 3, 14,071 were traced respondents in Cycle 4, 900 were traced non-respondents in Cycle 4, and 189 were not traced in Cycle 4.

## 9. Weighting

This chapter describes the weighting procedures for each subset of respondents described in Section 7.5. The longitudinal weighting process is necessarily different from that of cross-sectional weighting, for several reasons. First, longitudinal weights must represent the probability of selection of the unit of analysis at the time of sample selection. Since the longitudinal sample was selected in 1994-1995, the weights must reflect the probability of selecting the individual in Cycle 1 and not in subsequent Cycles. In addition, as explained in Section 7.5, the definition of a longitudinal response is different from that of a cross-sectional response, necessitating different adjustments particular to each type of non-response. Analysts should always use the longitudinal weights made from the subsets of respondents described in Section 7.5. The longitudinal weights have been calculated specifically to represent the 1994-1995 target population. In Cycles 1, 2 and 3 both cross-sectional and longitudinal files were produced. The weights attributed to the panel members on the cross-sectional files were different from those of the longitudinal file, and did not represent the correct probability of selecting the longitudinal member.

For Cycle 4, four sets of weights, WT64LS, WT60LF, WT60LFE and WT60SLF have been constructed. Table 9.A shows the subsets of respondents and the corresponding sampling weights and flags.

**Table 9.A: Subsets of Respondents and Corresponding Sampling Weights and Flags**

<b>Subsets of respondents</b>	<b>Sampling Weights</b>	<b>Flags</b>
Longitudinal Square	WT64LS	None, all records
Longitudinal Full	WT60LF	WF60LF
Longitudinal Full C1 and C4	WT60LFE	WF60LFE
Longitudinal Full Share	WT60SLF	WF60SLF

The WT60LF, WT60LFE and WT60SLF longitudinal weights have been adjusted for non-response and all four were post-stratified to the 1994-1995 population estimates based on 1996 Census counts by age group and sex within each province. Post-stratification is used to ensure that the four subsets of respondents represent correctly the 1994-1995 NPHS target population. The next section describes the NPHS longitudinal weighting method.

### 9.1 Longitudinal Weighting

The longitudinal weighting procedure is based on the weighting done for the Cycle 1 NPHS cross-sectional sample. Some weight adjustments that were applied to the Cycle 1 cross-sectional weights are no longer necessary for the subsequent longitudinal cycles. These adjustments were removed to create a “stripped” weight for each longitudinal panel member. This stripped weight is the starting point to obtain the longitudinal weights.

### **9.1.1 Cycle 1 (1994-1995) Stripped Weights**

The Cycle 1 stripped weights were obtained using the LFS basic weights as a starting point for all provinces except Quebec, where the basic weights from the “Enquête Sociale et de Santé” were taken as a starting point. Several adjustments were made to these weights to take into account the nature of the NPHS and to accurately represent the true probability of selection for each panel member. All of the adjustments that were made in Cycle 1 are kept for the subsequent cycles since the longitudinal sample always refers to the same population, that is the population of 1994-1995.

A full description of the Cycle 1 weighting procedures still relevant for subsequent cycles is included in sections 11.3 and 11.4 of the PUMF documentation for Cycle 2 and Cycle 3.

From this point, adjustments were made to the stripped weight to obtain the various sets of longitudinal weights.

### **9.1.2 Longitudinal Weights**

#### **9.1.2.1 Longitudinal Square Weight (WT64LS)**

The longitudinal square weight **WT64LS** is to be used with the square subset. It is calculated by post-stratifying the Cycle 1 stripped weight to the 1994-1995 population estimates based on 1996 Census counts by age group (0-11, 12-24, 25-44, 45-64, 65 and older) and sex within each province. The post-stratification adjustment is given by:

$$\frac{\text{Population estimate in a province/age/sex category}}{\text{Sum of weights of respondent household members in a province/age/sex category}}$$

#### **9.1.2.2 Longitudinal Full Weight (WT60LF)**

The longitudinal full subset includes only selected members who responded in all cycles or who have died. The weights of all non-respondents must be redistributed to compensate for the non-response.

The Cycle 1 stripped weight is the starting point and adjustments for non-response are made. A different non-response adjustment is made for each cycle, and these adjustments are cumulative from one cycle to another. For example, to obtain the Cycle 4 weights, the non-response adjustments for Cycles 2, 3 and 4 are applied successively to the Cycle 1 stripped weights.

The adjustments necessary in order to obtain the Longitudinal Full weight are described below.

***Adjustment 1: Adjustment for Cycle 2 (1996-1997) Non-response***

Adjusting for non-response was done using the weighting class approach. Weighting classes consist of groupings of respondents that share the same propensity to respond to the survey. Characteristics from Cycle 1, available for Cycle 2 respondents and non-respondents 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 non-response rate (and within which the non-response 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 chi-square automatic interaction detection (CHAID) algorithm was used to determine the weighting classes. In order to produce more stable adjustments, a minimum of 30 units per weighting classes was used.

Separate weighting classes were created for each province. Note that the province here refers to the province of residence at the time of the sample selection in 1994-1995. The Cycle 1 characteristics of the household as well as personal characteristics of the longitudinal member were considered. Some characteristics related to the sampling design of the survey or to the sampling weight were also considered in an effort to incorporate the sampling design of the survey into the analysis. Personal characteristics from the Health component were not used because they were not available for many longitudinal members in 1994-1995.

The variables chosen by the CHAID algorithm to build weighting classes to adjust for Cycle 2 non-response are listed in Table 9.B. Two variables from Cycle 1 sample design, one representing a flag which indicates the presence of members under 25 years old in the household and the other which indicates the presence of members under 12 years old in the household were used. The Cycle 1 non-response flag for income and the flag that indicates if the individual was under age 16 were also used. Please refer to the Data Dictionary for a complete description of the variables listed In Table 9.B.

**Table 9.B: Variables for Cycle 2 Non-response adjustment - WT60LF**

DHC4 AGE	DHC4 MAR	GE34DURB	LFC4 1	SDC4DRAC
DHC4DECf	DHC4 OWN	HCC4DMDC	RAC4F1	SDC4GCB7
DHC4 DWE	GE34DCMA	INC4DIA5	SDC4DAIM	SEX

To adjust for longitudinal members who did not respond in Cycle 2, the following adjustment is applied to the weight of respondents:

$$\frac{\text{Sum of weights for all longitudinal members}}{\text{Sum of weights for Cycles 1 and 2 responding longitudinal members}}$$

This adjustment is performed at the weighting class level.

***Adjustment 2: Adjustment for Cycle 3 (1998-1999) Non-response***

The 15,670 records with a full response after two cycles are taken as the starting point. The longitudinal pattern has been designed so that each year we simply add on a digit at the end of the variable to indicate the status of that year. In any given year, the code used in the construction of the pattern variable reflects the “state” of the panel member for that particular year. The codes are: 1=complete, 2=deceased, 3=institutionalized, 4=partial and 5=nonresponse. A “Full Longitudinal Response” after three cycles is defined as one of the following response patterns: 111, 112, 113, 122, 131, 132 or 133. All other response patterns are considered as non-responses (i.e., 114, 115, 134 and 135). Records for which the panel member was deceased in Cycle 2 (i.e., pattern 122) or institutionalized since Cycle 2 (i.e., pattern 133) are treated differently from the rest. For these records, no non-response adjustment is made since their weight in Cycle 2 has been already adjusted to reflect the fact that some of the Cycle 2 non-respondents may have in fact been deceased or institutionalized.

Adjusting for non-response was done using the weighting class approach. Separate weighting classes were created for each province (i.e., the 1994-1995 residence province). When adjusting for non-response in Cycle 3, only the Cycle 2 characteristics of the household as well as personal characteristics of the longitudinal member were considered. Again, as for Cycle 2, characteristics related to the sampling design of the survey or to the sampling weight were considered in an effort to incorporate the sampling design of the survey into the analysis. However, unlike for the Cycle 2 non-response adjustment, personal characteristics from the Health component were used, because they were available for all records that went into the Cycle 3 non-response adjustment.



The variables chosen by the CHAID algorithm to build weighting classes to adjust for Cycle 2 non-response are listed in table 9.C. The Cycle 1 sample design variable that represents an “Adult/Children” household type classification has also been used, as well as a Cycle 2 item non-response flag for income. Please refer to the Data Dictionary for a complete description of the variables listed in Table 9.C.

**Table 9.C: Variables for Cycle 3 / Non-response Adjustment - WT60LF**

AD_6_1	DHC6_AGE	INS6_4	SDC6DAIM	SMS6_9A
AD_6_7	DHC6_MAR	INS6_6	SDC6_4P	SMS6_13A
ALC6WKY	DV_6_65J	LFC6_41	SDC6_5A	SMS6_13C
ALC6_3	EDC6_3	MHC6DWK	SDC6_5F	SMS6_13 <sup>E</sup>
AM56_SHA	ES_6_80	MHC6_1A	SDC6_6B	SMS6_16D
AM66_PXY	GE36LMOV	MHC6_1B	SDC6_7A	SMS6_18A
AM66_SHA	HCC6F1	MHC6_1F	SDC6_7B	SMS6_18D
BPC6_10	HSC6DPAD	MHC6_1L	SDC6_7D	SP36_CPA
CCC6DNUM	HWS_5	MHC6_13	SEX	SSC6D2
CCC6_1L	INC6DIA5	PC_6_40	SHS6_4	SSC6_3
CCC6_1N	INC6_1A	RPC6_3	SMC6_2	SSS6_2
DGC6_1D	INC6_3B	RSS6_1	SMC6_5	SSS6_4

To adjust for longitudinal members that did not respond in Cycle 3, the following adjustment is applied to the weight of respondents:

$$\frac{\text{Sum of weights for Cycles 1 and 2 responding longitudinal members}}{\text{Sum of weights for Cycles 1, 2 and 3 responding longitudinal members}}$$

This adjustment is performed at the weighting class level, and is calculated from records with the following longitudinal response patterns: 111 to 115, 131, 132, 134 and 135. Again, records for which the panel member was deceased in Cycle 2 or institutionalized since Cycle 2 are not part of this adjustment.

**Adjustment 3: Adjustment for Cycle 4 (2000-2001) Non-response**

The 14,619 records with a full response after three cycles are taken as the starting point. Once again, records for which the panel member was deceased in Cycle 2 or 3 or institutionalized since Cycle 2 or 3 are treated differently from the rest. For these records, no non-response adjustment is made since their weight in Cycle 2 or 3 has been already adjusted to reflect the fact that some of the Cycle 2 or Cycle 3 non-respondents may have in fact been deceased or institutionalized.

Here again, adjusting for non-response was done using the weighting class approach. Separate weighting classes were created for each design province i.e. the 1994-1995 province of residence. When adjusting for non-response

in Cycle 4, only the Cycle 3 characteristics of the household as well as personal characteristics of the longitudinal member were considered. As for Cycle 3, characteristics related to the sampling design of the survey or to the sampling weight were considered in an effort to incorporate the sampling design of the survey into the analysis. Personal characteristics from the Health component were used, because they were available for all records that went into the Cycle 4 non-response adjustment.

The variables chosen by the CHAID algorithm to build the weighting classes to adjust for Cycle 4 non-response are in Table 9.D. A Cycle 3 item non-response flag for income has also been used. Please refer to the Data Dictionary for a complete description of the variables listed in table 9.D.

**Table 9.D: Variables for Cycle 4 / Non-response Adjustment - WT60LF**

CCC8DANY	DGC8_1A	HCC8_1	PAC8_1A	SDC8_6A
CCC8_1C	DHC8_AGE	INC8DIA5	PAC8_1J	SDC8_7A
CCC8_1L	DHC8DECF	ISC8_1	PY_8DH1	SEX
CCC8_1N	DHC8_OWN	NU_8_1B	RAC8F1	SSC8DEMO
CCC8_1V	FIC8F1	PAC8DFD	RPC8_2	SSC8DSOC
DGC8F1	GE38DURB	PAC8DLEI	SDC8_4A	TWC8_5

To adjust for longitudinal members that did not respond in Cycle 4, the following adjustment is applied to the weight of respondents:

$$\frac{\text{Sum of weights for Cycles 1, 2 and 3 responding longitudinal members}}{\text{Sum of weights for Cycles 1 to 4 responding longitudinal members}}$$

This adjustment is performed at the weighting class level. Records for which the panel member was deceased in Cycle 2 or 3 or institutionalized since Cycle 2 or 3 are not part of this adjustment.

**Adjustment 4: Post-stratification adjustment**

The weight of responding units has been post-stratified to the 1994-1995 population estimates based on 1996 Census counts by age group (0-11, 12-24, 25-44, 45-64, 65 and older) and sex within each province. This is done to ensure that the 1994-1995 population is accurately represented in any estimates produced from the longitudinal file. This adjustment is given by

$$\frac{\text{Population estimate in a province/age/sex category}}{\text{Sum of weights of Cycles 1 to 4 responding longitudinal members in a province/age/sex category}}$$

The final longitudinal weight **WT60LF** is calculated by taking the Cycle 1 stripped weight and multiplying that value by Adjustments 1 to 4.

### **9.1.2.3 Longitudinal Full Weight for Cycles 1 and 4 (WT60LFE)**

The NPHS now includes, starting with Cycle 4, an additional weight to be used with the Full C1 and C4 subset of respondents. This new subset includes the respondents that are part of the full subset, as well as those respondents with a full response in Cycles 1 and 4 only (that is, some non-respondents and partial respondents in Cycle 2 or Cycle 3). Since a large number of the respondents in this subset are also part of the full subset, the method used for the weighting of the full subset is once again used, with a few modifications.

The starting point is the Cycle 4 longitudinal full weight, just before the post-stratification adjustment. As described in previous sections, this weight has been adjusted for non-response to each cycle. However, some Cycle 2 or Cycle 3 non-respondents whose weights had been distributed to the respondents that make up the full subset were respondents once again in Cycle 4 and their weights now have to be retrieved. The starting point for each retrieved record in Cycle 4 is its weight from the most recent cycle where it was part of the full subset. This allows the use of the most recent weight that takes the non-response adjustment into account for previous cycles. An amount equivalent to these distributed weights must be removed from the Cycle 4 full subset. This removal is done separately for each of the weighting classes in the cycle where the retrieved records were non-respondents for the first time. It involved giving the weights back to the records in the same way that they were lost. However, this weight must be adjusted so that the records retrieved in Cycle 4 represent the same proportion in the new full response in Cycles 1 and 4 subset as they would represent if they had been part of the full subset for each cycle, that is, if they had been involved in all of the non-response adjustments for subsequent cycles. This adjustment is done separately within each weighting class.

Finally, as for the other longitudinal weights, the weights of those records belonging to the new subset were corrected by post-stratifying to 1994-1995 population estimates based on the 1996 Census counts by age group (0-11, 12-24, 25-44, 45-64, 65 and over) and sex within each province. This adjustment is given by :

$$\frac{\text{Population estimate in a province/age/sex category}}{\text{Sum of weights of the Full C1 and C4 subset members in a province/age/sex category}}$$

The final weight for a full response in Cycles 1 and 4 is the weight WT60LFE. A technical document providing more details on the weighting method for the full response in Cycles 1 and 4 subset will eventually be available on request.

#### **9.1.2.4 Longitudinal Full Share Weight (WT60SLF)**

In some cases, respondents indicated that they did not want to share the information provided from all interviews conducted as part of NPHS with provincial ministries of health and Health Canada. As these partners only receive the records of these sharers, a special weight must be derived so that estimates computed from this subset correctly represent the total population.

A simple adjustment is made to the longitudinal full weight to create the share weight. This adjustment is given by:

$$\frac{\text{Sum of weights for Cycles 1 to 4 responding longitudinal members in a province / longitudinal pattern / age-sex category}}{\text{Sum of weights for Cycles 1 to 4 responding longitudinal members who agreed to share, in a province / longitudinal pattern / age-sex category}}$$

Note that in Cycles 3 and 4, a few of the original longitudinal response patterns were collapsed in order to produce more stable adjustments. The grouping was done for a few province/age-sex categories that had few observations in some of the longitudinal patterns representing deceased or institutionalized. In each case, the problematic response pattern was grouped with another longitudinal pattern in the same province/age-sex category, so that the sum of the weights would still give the correct population counts. The collapsing of longitudinal response patterns preserved the weighted distribution of panel members living in households, institutionalized or deceased for each of these cycles. The final longitudinal sharing weight (WT60SL) is obtained by multiplying the longitudinal full weight (WT60LF) by this adjustment. Note that since this adjustment is done with respect to the post-stratification classes, no additional post-stratification is necessary.

## 10. Data Quality

### 10.1 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.

#### 10.1.1 Non-sampling Errors

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

A major source of non-sampling errors in surveys is the effect of non-response on the survey results. The extent of non-response varies from partial non-response (failure to answer just one or some questions) to total non-response. Partial non-response to NPHS was minimal; once the questionnaire was started, it tended to be completed with very little non-response. 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. Total non-response occurred because it was impossible 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 persons who responded to the survey to compensate for those who did not respond. See Section 9.1.2.2 for details of the weight adjustment for non-response.

### **10.1.2 Sampling Errors**

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

For example, suppose hypothetically that one estimates that 25% of Canadians aged 12 and over have experienced an improvement in their general health between Cycle 1 and Cycle 2 of the survey and that this estimate is found to have a standard deviation of .003. Then the CV of the estimate is calculated as:

$$(.003/.25) \times 100\% = 1.20\%.$$

Statistics Canada commonly uses CV results when analyzing data, and urges users producing estimates from NPHS data files to also do so. For guidelines on how to interpret CV results, see the table at the end of Section 11.4.

#### **10.1.2.1 Bootstrap Method for Variance Estimation**

In order to determine the quality of the estimate and to calculate the CV, the standard deviation must be calculated. Confidence intervals also require the standard deviation of the estimate.

The NPHS uses a multi-stage survey design, which means that there is no simple formula that can be used to calculate variance estimates. Therefore, an approximative method was needed. The bootstrap method is used because the sample design information needs to be taken into account when calculating variance estimates. The bootstrap method does this, and with the use of the Bootvar program, discussed in the next subsection, remains a method that is fairly easy for users to use.

The bootstrap re-sampling method used in the NPHS involves the selection of simple random samples known as replicates, and the calculation of the variation in the estimates from replicate to replicate. In each stratum, a simple random sample of (n-1) of the n clusters is selected with replacement to form a replicate. Note that since the selection is with replacement, a cluster may be chosen more than once. In each replicate, the survey weight for each record in the (n-1) selected clusters is recalculated. These weights are then post-stratified according to demographic information in the same way as the sampling design weights in order to obtain the final bootstrap weights.

The entire process (selecting simple random samples, recalculating and post-stratifying weights for each stratum) is repeated B times, where B is large. The NPHS typically uses B=500, to produce 500 bootstrap weights. To obtain the bootstrap variance estimator, the point estimate for each of the B samples must be calculated. The standard deviation of these estimates is the bootstrap variance estimator. Statistics Canada has developed a program that can perform all of these calculations for the user: the Bootvar program.

#### **10.1.2.2 Bootvar Program for Variance Estimation**

The Bootvar program is available in both SAS and SPSS formats. It is made up of macro that compute variances for totals, ratios, differences between ratios and for linear and logistic regression.

The Bootvar program is provided with bootstrap weights and a document explaining how to modify and use the program to suit user's needs.

## **11. Guidelines For Tabulation, Analysis and Release**

This section of the documentation outlines the guidelines that should be followed by users to tabulate, analyze, release or otherwise publish any data derived from the NPHS data. With the aid of these guidelines, users 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.

### **11.1 Rounding Guidelines**

In order that dissemination of estimates derived from NPHS data corresponds to estimates produced by Statistics Canada, Users should use the following guidelines regarding the rounding of such estimates. Unrounded estimates imply greater precision than actually exists.

- a) Estimates in the main body of a statistical table should 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 preceding digit is incremented by 1.
- b) Marginal sub-totals and totals in statistical tables should 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 should 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) should 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, it is suggested to users to note the reason for such differences in the publication or release document(s).



## **11.2 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 statistical tables, users must apply the proper sampling weight. If proper weights are not used, the estimates derived from the various subsets of respondents cannot be considered to be representative of the 1994-1995 target population, and will not correspond to those produced by Statistics Canada.

Many software packages compute estimates using sampling weights from complex surveys. However, these software packages use various methods to include the sampling weights in their calculations of estimates. Users should note that some software packages might not generate estimates that exactly match those available from Statistics Canada.

### **11.2.1 Definitions of Types of Estimates: Categorical vs. Quantitative**

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 computed.

#### ***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 quit smoking between cycles 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.

#### ***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 increase in the number of cigarettes smoked per day by daily smokers who had an increase in consumption between two cycles.

### **11.2.2 Tabulation of Categorical Estimates**

Estimates of the number of people with a certain characteristic can be obtained by summing the final weights of all records possessing the characteristic(s) 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.

### **11.2.3 Tabulation of Quantitative Estimates**

Estimates of quantities can be obtained from the 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.

## **11.3 Guidelines for Statistical Analysis**

The NPHS is based upon a complex sampling 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 that which 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 final weight divided by the average of the final weights for the sampled units (people) contributing to the estimate in question.

CV tables were produced in the past for the cross-sectional data. CV tables were not created for the longitudinal files as a very large number of possible variable combinations for analysis exist. To correctly estimate the variance, NPHS recommends the use of the bootstrap method. With the bootstrap method, the complexity of the weighting and the survey design are incorporated into the calculation of the variance. A SAS bootstrap variance program, along with accompanying documentation and examples of how to use it, has been created to facilitate the calculation of the variance using the bootstrap method. The program also calculates the accompanying coefficient of variation. A similar version of the program is also available in SPSS. It is important for users to learn how to use it as the

program will generate more exact estimates of individual variances to assess the quality of tabulated estimates and is highly recommended over the use of the scaled weights approach. Some statistical packages such as STATA have the ability to read in the stratum and cluster information to use in variance estimation, which improves the quality of the estimate but does not take into account the different adjustments applied to the weights.

**11.4 Release Guidelines**

Before releasing and/or publishing any estimate from the master files, users should first determine the number of sampled respondents who contribute to the calculation of the estimate. If this number is less than 10, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. This is due to the fact that the possibility of obtaining an artificially low variance is greater with a sample size less than 10. For weighted estimates based on sample sizes of 10 or more, users should determine the coefficient of variation of the estimate and follow the guidelines described in Table 11.A.

**Table 11.A: Sampling Variability Guideline**

<b>Type of Estimate</b>	<b>C.V. (in %)</b>	<b>Guidelines</b>
Acceptable	0.0 - 16.5	Estimates can be considered for general unrestricted release. Requires no special notation.
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 E (or in some other similar fashion).
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 F (or in some other fashion) and the following warning should accompany the estimates:  The user is advised that . . .(specify the data) . . . do not meet Statistics Canada's quality standards for this statistical program. Conclusions based on these data will be unreliable and most likely invalid. These data and any consequent findings should not be published. If the user chooses to publish these data or findings, then this disclaimer must be published with the data.

## **12. Using the Longitudinal Master Files**

### **12.1 Use of Longitudinal Weights**

In past cycles, a few files made up of subsets of the 17,276 longitudinal panel members were created. This time, only one file has been produced. The Cycle 4 master file contains four subsets of respondents (Section 7.5) to which correspond a set of weights (Section 9.1). Flags were created to identify records that are part of a particular subset (Table 9.A). Records that are not part of a particular subset have a flag equal to 0 and the weight variable set to blank for that particular subset. To create the subset of interest, select those records that have the appropriate flag variable equal to 1.

Weight WT64LS is called the “square weight” and applies to the “Square” subset of respondents which includes all 17,276 members that make up the original longitudinal sample. This weight is to be used with Users who wish to do specialized studies on non-response bias.

Weight WT60LF is called the “Longitudinal Full” weight and applies to the 13,582 records that are included in the “Full” subset of respondents.

Weight WT60LFE is called the “Longitudinal Full C1 and C4” weight and applies to the 14,321 records that are included in the “Full C1 and C4” subset of respondents.

Weight WT60SLF is called “Longitudinal Full Share” weight and applies to the 13,110 respondents that are included in the “Full Share” subset of respondents.

### **12.2 Ensuring the Reliability of Estimates with the Use of Bootstrap Weights**

Bootstrap weights are necessary for variance estimation. Information on the bootstrap method for variance estimation can be found in Section 10.1.2.1. Each subset of respondents has a set of bootstrap weights associated with it. Four different sets of bootstrap weights were created for the Cycle 4 data: the square, the full, the full share and the full C1 and C4. For more information on these subsets, see Section 7.5. Table 12.A presents the subset of respondents with their corresponding bootstrap file name.

**Tableau 12.A: Subsets of Respondents and Corresponding Bootstrap Weights files**

<b>Subset of respondents</b>	<b>Bootstrap Weights file</b>
Longitudinal Square	B5long
Longitudinal Full	B5lngf
Longitudinal Full C1 and C4	B5lngfe
Longitudinal Full Share	B5lngf (share)

Due to the complex sample design, users should use the Bootvar program for variance calculation. The standard variance output from other statistical packages such as SAS and SPSS may grossly underestimate the variance of an estimate for this survey. **It is the responsibility of the user to ensure the quality/reliability of the estimates that they are producing by following the guidelines laid out in Chapter 11 and correctly calculating the variance for all estimates.** Failure to do so could lead to some misinterpretation of results and jeopardize the quality of the research work.

Some statistical software are capable of including the stratum and cluster information as input when performing analytical processing, which does provide a variance estimate much closer to the true variance estimate, but these packages fail to account for the various weighting adjustments, which in some cases can impact the variance estimates considerably.

### **12.3 Variable Naming Convention**

NPHS has 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 in 1998-1999, and the variable names should only differ in the position that identifies the particular survey occasion in which they were collected. This convention is followed throughout the longitudinal survey, and is adopted by all NPHS surveys: the household component, the health care institutions component, and previously the North component and supplements.

### 12.3.1 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 name / 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, DHC8\_AGE, and DHC0\_AGE:

**DH:** in the Demographic and Household content section of the questionnaire;  
**C:** questions which are Core content on the household survey;  
**4/6/8/0:** appeared in Cycle 1 (1994-1995), / appeared in Cycle 2 (1996-1997), / appeared in Cycle 3 (1998-1999), / appeared in Cycle 4 (2000-2001);  
**\_:** can be found on the questionnaire;  
**AGE:** the variable name.

### 12.3.2 Positions 1-2: Variable name / Questionnaire Section Name

AD	Alcohol dependence	IS	Insurance
AL	Alcohol	LF/LS	Labour force
AM	Administration of the survey	MH	Mental health
AP	Attitudes towards parents	NU	Nutrition
BP	Blood pressure	PA	Physical activities
CC	Chronic conditions	PC	Physical check-up
CE	Contact/exit	PH	Preventive health
CO	Coping (Alberta buy-in, Cycles 1 and 2)	PR	Province
DG	Drug use	PY	Psychological resources (self-esteem, mastery, sense of coherence)
DH	Demographics and household	RA	Restriction of activities
DV	Dental visits	RP	Repetitive strain
ED	Education	RS	Road safety
ES	Emergency services	RT	Rationality (Manitoba buy-in, Cycle 1)
EX	Eye examination	SC	Self-care
FH	Family history	SD	Socio-demographics

***NPHS Cycle 4 (2000-2001), Longitudinal Documentation***

FI	Food Insecurity (Human Resources Development Canada, Cycle 3 Buy-in)	SH	Sexual health
FS	Flu shots	SM	Smoking
GE	Geographic identifiers	SP	Sample control variables
GH	General health	SS	Social support
HC	Health care utilization	ST	Stress
HI	Health information	SV	Health services
HN	Health number	TA	Tobacco alternatives
HS	Health status	TU	Tanning and UV exposure
HV	HIV	TW	Two-week disability
HW	Height and weight	VS	Violence / personal safety
IJ	Injuries	WH	Women's health: breast self-examination, breast examination, mammography, Pap smear, and hysterectomy
IN	Income	WF	Subset flags
		WT	Subset sampling weights

A few important identifying variables do not follow the naming convention: e.g. REALUKEY, PERSONID, CYCLE, SUBCYCLE, DESIGPRV, STRATUM, and REPLICAT.

There are also some variables that are considered “constant”. Table 12.B presents the variables that appear only once of the data file. The name of these variables does not follow the naming convention.

**Table 12.B: “Constant” Longitudinal Variables**

<b>Longitudinal Name</b>	<b>Concept</b>
SEX	Sex
DOB	Day of birth
MOB	Month of birth
YOB	Year of birth
COB	Country of birth
COBC	Code of country of birth
COBGC	Country of birth (7 groups) - grouped
IMM	Flag indicating that the respondent is an immigrant
AOI	Age at time of immigration - derived
HWB	Birth weight
DOD	Day of death
MOD	Month of death
YOD	Year of death
COD	Cause of death code

**12.3.3 Position 3: Survey Type**

- A Asthma supplement
- B Province-specific buy-in content – children’s questions
- C: Core questions repeated in each cycle
- F: Food Insecurity supplement
- 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)
- \_ : Cycle specific focus questions, not repeated in every cycle (e.g., stress in Cycles 1 and 4, access to services in Cycle 2)
- 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)

**12.3.4 Position 4: Year / Cycle Variable**

- 4 Cycle 1 (1994-1995)
- 6 Cycle 2 (1996-1997)
- 8 Cycle 3 (1998-1999)
- 0 Cycle 4 (2000-2001)
- 2 Cycle 5 (2002-2003)
- A Cycle 6 (2004-2005)
- B Cycle 7 (2006-2007)
- C Cycle 8 (2008-2009)
- D Cycle 9 (2010-2011)
- E Cycle 10 (2012-2013)
- F Cycle 11 (2014-2015)



**12.3.5 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., North American Industry Classification System (NAICS))
D	Derived variable	A variable calculated from one or more collected or coded variables, usually calculated during head office processing (e.g., Comprehensive Health Status Measurement System (CHSMS-HUI3))
F	Flag variable	A variable calculated from one or more collected variables (like a derived variable), but usually calculated by the computer application for later use during the interview (e.g., work flag). It can also denote that a long answer was collected (e.g., restriction of activity 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

**12.3.6 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 DHC4\_AGE in Cycle 1 (1994-1995), DHC6\_AGE in Cycle 2 (1996-1997), etc. 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. 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.

## 13. Access to NPHS Data

### 13.1 Microdata Files

Confidentiality concerns preclude general dissemination of longitudinal NPHS data in public use microdata file (PUMF) format. However, access to the longitudinal master microdata files including the Cycle 4 data (as well as access to the cross-sectional master microdata files, which exist for the first three cycles of the NPHS) is available through Health Statistics Division's Remote Access service. This service provides researchers with a means to develop and test their own computer programs using synthetic files that mimic the actual master files. Researchers then submit their programs to a dedicated e-mail address. The programs are run against the master microdata files on an internal secure server, outputs are vetted for confidentiality, and sent back to the researcher by return e-mail. For more information on this service, please contact the Data Access team at [nphs-ensp@statcan.ca](mailto:nphs-ensp@statcan.ca).

Direct, on-site access to the NPHS master microdata files is also possible at Statistics Canada's Research Data Centres (RDCs). These centres, established in collaboration with the Social Sciences and Humanities Research Council (SSHRC), are situated in secure physical locations at participating universities. They operate as extensions of Statistics Canada offices, with a full-time Statistics Canada employee at each centre, and researchers conduct their work under the terms of the *Statistics Act*, as would any other Statistics Canada employee. More information is available at the Social Sciences and Humanities Research Council Web site: <http://www.sshrc.ca/rdc>.

PUMFs are available for each of the first three cycles of the NPHS, providing widespread access to the cross-sectional components of the survey. The NPHS PUMFs can be accessed through the Data Liberation Initiative (DLI) at participating Canadian universities and colleges. For more information, please consult Statistics Canada's Web site at <http://www.statcan.ca/english/edu/index.htm>. Cycles 1, 2 and 3 NPHS PUMFs can also be purchased. To this end, please contact Health Statistics Division's technical support team at [hd-ds@statcan.ca](mailto:hd-ds@statcan.ca) or one of Statistics Canada's Regional Offices.

### 13.2 Analytical Reports and Tabulations

Research articles based on the NPHS often appear in Health Reports, a quarterly journal produced by Health Statistics Division. This product is available as a standard printed publication (catalogue no. 82-003-XPE) or in electronic format on the Statistics Canada Internet site as catalogue no. 82-003-XIE. To obtain more information, visit our Web site at [www.statcan.ca](http://www.statcan.ca), and select Products and Services.

Custom tabulations from the NPHS are available on a cost recovery basis. For estimates on costs and feasibility, contact the Health Statistics Division's technical support team at [hd-ds@statcan.ca](mailto:hd-ds@statcan.ca).

Finally, the Guide for Health Statistics (82-573-GIE) on Statistics Canada's Web site is a good starting point to health-related information with links to health indicators from various sources including the NPHS. The Guide also links to various documents related to the NPHS, including the questionnaires. Visit the Guide at the following address: <http://www.statcan.ca/english/freepub/82-573-GIE/guide.htm>.

**Appendix A – NPHS Household Component, Health Content over 4 cycles**

<b>NPHS Core Content: Cycle 1 to Cycle 4</b>				
	<b>Cycle 1 (1994-1995)</b>	<b>Cycle 2 (1996-1997)</b>	<b>Cycle 3 (1998-1999)</b>	<b>Cycle 4 (2000-2001)</b>
AL - Alcohol	√	√	√	√
AM - Administration	√	√	√	√
CC - Chronic Conditions	√	√	√	√
DG - Drug Use	√	√	√	√
DH - Demographics and Household	√	√	√	√
ED - Education	√	√	√	√
GH - General Health	√	√	√	√
HC - Health Care Utilization	√	√	√	√
HS - Health Status	√	√	√	√
HW - Height / Weight	√	√	√	√
IJ - Injuries	√	√	√	√
IN - Income	√	√	√	√
IS - Insurance		√	√	√
LF / LS - Labour Force	√	√	√	√
MH - Mental Health	√	√	√	√
PA - Physical Activities	√	√	√	√
PH - Preventive Health	√		√	√
RA - Restriction of Activities	√	√	√	√
RP - Repetitive Strain		√	√	√
SD - Socio-demographic Characteristics	√	√	√	√
SM - Smoking	√	√	√	√
SS - Social Support	√	√	√	√
TU - Tanning and UV Exposure				√
TW - Two-week Disability	√	√	√	

***NPHS Cycle 4 (2000-2001), Longitudinal Documentation***

---

<b>NPHS Focus Content: Cycle 1 to Cycle 4</b>				
	<b>Cycle 1 (1994-1995)</b>	<b>Cycle 2 (1996-1997)</b>	<b>Cycle 3 (1998-1999)</b>	<b>Cycle 4 (2000-2001)</b>
AD - Alcohol Dependence		√		
BP - Blood Pressure		√		
CC - Chronic Conditions (arthritis, heart disease, diabetes)				√
DV - Dental Visits		√		
ES - Emergency Services		√		
EX - Eye Examinations		√		
FH - Family Medical History			√	
FS - Flu Shots		√		
GH – Breast-feeding		√		
NU - Nutrition			√	
PC - Physical Check-up		√		
PY - Sense of Coherence	√		√	
SC - Self Care			√	
ST - Stress (ongoing problems, recent life events, childhood and adult stressors "traumas", work stress, self-esteem, mastery)	√			√
WH - Preventive Health (breast examinations, breast self-examinations, mammography, Pap Smear Test)		√		

***NPHS Cycle 4 (2000-2001), Longitudinal Documentation***

<b>NPHS Buy-in Content: Cycle 1 to Cycle 3</b>			
	<b>Cycle 1 (1994-1995)</b>	<b>Cycle 2 (1996-1997)</b>	<b>Cycle 3 (1998-1999)</b>
Health Promotion Survey		√	
AL - Alcohol		√	
FI - Food Insecurity		√	√
GH - Breast-feeding		√	
GH - Pregnancy		√	
HW - Height / Weight		√	
HV - HIV		√	
NU - Diet / Nutrition		√	√
RS - Road Safety		√	
SH - Sexual Health		√	
SM - Smoking		√	
TA - Tobacco alternatives			√
WH - Breast Self-examination		√	
Provincial Content Buy-in	√	√	
AP - Attitudes towards Parents		√	
CO - Coping (Manitoba)	√		
CO - Coping (Alberta)	√	√	
HI - Health Information		√	
SH - Sexual Health		√	
SS - Social Support		√	
SV - Health Services		√	
SV - Child Health Services		√	
TU - Tanning and UV Exposure		√	
VS - Violence and Personal Safety		√	

**Appendix B: NPHS Household Component, Changes to the questionnaire for Cycle 4 (2000-2001)**

**Changes to core content**

In the following description, external variable names from the questionnaires are used. External variable names for the master and share files are based on the variable naming convention (see Section 12.3).

**AM - Change of Residence**

- Deletion: Questions on usual place of residence in Cycle 1 (AMC8\_MV1, AMC8\_MV2, and AMC8\_MV3) - solely longitudinal survey.

**AL – Alcohol**

- Addition: Question on whether person drank alcohol since last interview (ALC0\_1A) - If the person answers “no” this creates a new skip pattern around the section. If the person says “yes” then he/she is asked whether they have had a drink in the past 12 months (ALC0\_1). If the person says “no” then he/she is not asked if he/she has ever had a drink, since (ALC0\_1A) is “yes”.

**AM – Administration**

- Deletion: Question on agree to share information (AM60\_SHA) “Human Resources Development Canada (HRDC)” dropped as listed share organization.
- Addition: Question on having a health number to help in linkage (AM60\_H3A) .
- Addition: Question on province of health care number (AM60\_H3B).
- Addition: Validity edit algorithms for provincial health numbers AM\_HN - built into CAI for most provinces.

**CC - Chronic conditions**

- Addition: Question on fibromyalgia (CCC0\_1X).
- Change: Question on arthritis and rheumatism (CCC0\_1D) – “excluding fibromyalgia” added.
- Change: Question on back problems (CCC0\_1E) - “excluding fibromyalgia” added.
- Addition: Questions on age at first diagnosis for each chronic condition except “other” (CCC0\_A3 to CCC0\_X3).
- Change: Question on currently take insulin for diabetes (CCC0\_J5) - wording clarification added “currently” and “for this” replaced by “for your diabetes”.
- Change: Question on take any other treatment or medication for diabetes (CCC0\_J6) - wording clarification “for this” replaced by “for your diabetes”.
- Deletion: Questions on chronic nose, throat, and ear infections for children less than 3 years of age (CCK8\_1, CCK8\_2, and CCK8\_3).
- Deletion: Question on sinusitis (CCC8\_1I).

- Deletion: Questions on month and year of diagnosis for selected conditions (CCC8\_C3M to CCC8\_O3M, and CCC8\_C3Y to CCC8\_O3Y) - replaced by questions on age at first diagnosis.
- Deletion: Questions on confirmation of condition existing before previous interview for selected conditions (CCC8\_C4 to CCC8\_O4) - deleted because month and year of diagnosis not asked.
- Deletion: Questions on treatment for arthritis / rheumatism (CCC8\_D5) and mark-all kind of treatment or medication (CCC8\_D6A to CCC8\_D6D) - replaced with similarly worded focus questions (CC\_0\_D3 and CC\_0\_D3A to CC\_0\_D3H).
- Integration: Focus content on diagnosis and management - see section below on changes to focus content.

### **DG - Drug use**

- Change: Question on examples of diet pills (DGC0\_1C) - “diet pills such as Redux, Ponderal or Fastin” added.
- Change: Question on examples of anti-depressants (DGC0\_1D) - “anti-depressants such as Prozac, Paxil or Effexor” added.
- Addition: Question on use of other health products in last 2 days (DGC0\_4A).
- Addition: Question on specific name of other health product used in last 2 days (DGC0C5A) and (DGC0C5B to DGC0C5L) - up to 12 products.
- Deletion: Question on specific name of other health product used (no time frame specified) (DGC8C5A to DGC8C5L) - up to 12 products.

### **ED – Education**

- No change.

### **GH - General health**

- Addition: Question on self-perceived level of stress (STC0\_1).

### **HC - Health care utilisation**

- Deletion: Questions on place of most recent contact with family doctor (HCC8\_3A), other medical doctor (HCC8\_3C) and nurse (HCC8\_3D) - included in CCHS.
- Deletion: Questions on health care received in the United States (HCC8\_11 and HCC8\_12) – included in CCHS.
- Addition: Question on having a regular medical doctor (TWC0\_5).

### **HS - Health status**

- No change.

### **HW - Height and weight**

- Deletion: Question on weight at birth (GHK8\_6) - no new babies in sample.



## **IJ - Injuries**

### **Major revision to include new content and bring categories in line with ICD-10**

- Revision: Question on type of injury, response categories (IJC0\_3).
- Addition / revision: Question on part of body injured, response categories (IJC0\_4).
- Addition: Question on part of body injured (IJC0\_4A) - detail for injury to internal organs.
- Addition / revision: Question on place of occurrence of injury, response categories (IJC0\_5).
- Addition / revision / deletion: Question on type of activity (IJC0\_9), injury result of a fall (IJC0\_10), type of fall (IJC0\_10A), and cause of injury (IJC0\_10B) – replaces (IJC8\_6).
- Addition: Question on received medical attention within 48 hours (IJC0\_11) .
- Addition: Question on treatment site (IJC0\_12).
- Addition: Question on admitted to hospital (IJC0\_13).
- Addition: Question on treated but non-limiting injuries (IJC0\_14).
- Addition: Question on number of treated but non-limiting injuries (IJC0\_15).
- Deletion: Question on work-related injury (IJC8\_7).
- Deletion: Question on mark all precautions to prevent re-injury (IJC8\_8A to IJC8\_8G).

## **IN – Income**

- No change.

## **IS – Insurance**

- No change.

## **LF / LS - Labour force**

### **Major revisions to include new content and bring concepts closer to Labour Force Survey**

In general terms, the concept of current status has been replaced by status last week. The roster of up to 3 jobs has been replaced by main job and all other jobs. The definition of main job has not changed. The start and stop dates for up to three jobs (giving the derived variables on gaps) have been replaced with the number of weeks working, looking for work, and not working in the last 12 months, plus a question on the number of periods searching. There are also many new questions such as job search and main reason for type of working hours.

- Addition / revision: Question on had a job last week (LSC0\_1) replaces currently working (LFC8\_61, LFC8\_62, and LFC8\_63) – wording revision “working for pay or profit” replaced by “work at a job or business”.
- Addition: Question on had job last week but was absent (LSC0\_2).
- Addition: Question on had more than one job last week (LSC0\_3) - replaces (LFC8\_111 and LFC8\_112).
- Addition: Question on did something to find work last 4 weeks (LSC0\_11).

- Addition: Question on had definite start date in future last week (LSC0\_12).
- Addition / revision: Question on reason not currently working (LSC0\_13) - new response categories, replaces (LFC8\_17B).
- Addition / revision: Question on worked in last 12 months (LSC0\_21) - replaces (LFC8\_2).
- Addition: Question on did something to find work in last 12 months (LSC0\_22).
- Addition: Question on had more than one job in last 12 months (LSC0\_23) - universe excludes those working in last week.
- Addition / revision / deletion: Question on employee or self-employed (LSC0\_31) - replaces (LFC8\_16), change in wording of response category from “unpaid family worker” to “working in a family business without pay”.
- Addition / revision: Question on self-employed “name of business” (LSC0F32 with LSC0F33) - replaces (LFC8FE1).
- Addition / revision: Question on employee “name of employer” (LSC0F33 with LSC0F32) - replaces (LFC8FE1).
- Addition / revision: Question on “kind of business” (LSC0F34) minor wording change - replaces (LFC8F13).
- Addition / revision: Question on “kind of work” (LSC0F35) minor wording change - replaces (LFC8F14).
- Addition / revision: Question on “duties” (LSC0F36) minor wording change - replaces (LFC8F15).
- Addition / revision: Question on reason for absence last week (LSC0\_41).
- Addition / revision: Question on number of hours of work (LSC0\_42) see also (LSC0\_53) - replaces (LFC8\_81, LFC8\_82, and LFC8\_83) on each job.
- Addition: Question on work / pay preference (LSC0\_43).
- Addition / revision: Question on type of working hours main job (LSC0\_44) - replaces (LFC8\_91, LFC8\_92, and LFC8\_93) on each job.
- Addition / revision: Question on main reason for type of working hours (LSC0\_45).
- Addition / revision: Question on usually works weekend main job (LSC0\_46) see also (LSC0\_54) - replaces (LFC8\_101, LFC8\_102, and LFC8\_103) on each job.
- Addition / revision: Question on weeks in the past 12 months with more than one job (LSC0\_51).
- Addition / revision: Question on main reason for working at more than one job (LSC0\_52).
- Addition / revision: Question on hours of work all other jobs (LSC0\_53) see also (LSC0\_42) - replaces (LFC8\_81, LFC8\_82, and LFC8\_83).
- Addition / revision: Question on usually works weekend all other jobs (LSC0\_54) see also (LSC0\_46) replaces (LFC8\_101, LFC8\_102, and LFC8\_103) on each job.
- Addition / revision: Question on number of weeks worked last 12 months (LSC0\_61).
- Addition / revision: Question on number of weeks looking for work last 12 months (LSC0\_71).

## ***NPHS Cycle 4 (2000-2001), Longitudinal Documentation***

---

- Addition / revision: Question on number of weeks not working nor looking for work last 12 months (LSC0\_72).
- Addition / revision: Question on main reason not looking for work (LSC0\_73).
- Addition / revision: Question on periods looking for work (LSC0\_74).
- Deletion: Question on start and stop dates for each job (LFC8\_71D, LFC8\_71M, LFC8\_71Y, LFC8\_72D, LFC8\_72M, LFC8\_72Y, LFC8\_73D, LFC8\_73M, LFC8\_73Y).
- Deletion: Question on confirmation of working for previous employer (LFC8\_2B).

### **MH - Mental health**

- No change.

### **PA - Physical activities**

- No change.

### **PH - Preventive health**

- Deletion / new skip pattern: Questions on confirmation of previous inconsistencies on blood pressure (BPC8\_10A), PAP smear (WHC8\_20A) and mammogram (WHC8\_30A) - these were deleted since the answer from Cycle 3 now represents the confirmed response.
- Change: Question on last mammogram (WHC0\_32) from “when was the last time?” to “when was the last time that you had a mammogram?” - clarification of wording.
- Addition: Question on reason for mammogram, a new response category was added “breast problem” (WHC0\_33H) - this question was last asked in Cycle 2.
- Addition / change: Question on reasons for hysterectomy (WHC0\_5BA to WHC0\_5BG) – type of question changed from a “mark one only” to a “mark all that apply”.

### **RA - Restriction of activities**

- No change.

### **RP - Repetitive strain**

- Addition / revision: Question on body part affected (RPC0\_3) - response categories expanded from 9 to 13 categories. Category list more closely resembles revised injury question on part of body injured (IJC0\_4).
- Addition: Question on mark all activity type when strained (RPC0\_5A to RPC0\_5F).
- Deletion: Question on mark all location / activity of strain (RPC8\_4A to RPC8\_4D).

### **SD - Socio-demographic characteristics**

- Deletion: Questions on country of birth (SDC8\_1) and year of immigration (SDC8\_3).

### **SM – Smoking**

- Addition: New skip pattern (SM\_C103) current daily smokers who were also previous daily smokers are no longer asked the age they began to smoke cigarettes daily - data previously collected.
- Addition: New skip pattern (SM\_C105D) previous daily smokers are no longer asked if they have ever smoked cigarettes daily - data previously collected.

### **SS - Social support**

- No change.

### **TU - Tanning and UV exposure**

- Addition: Question on sunburnt in last 12 months (TUC0\_3).

### **TW - Two-week disability**

- Deletion: Questions on bed days and cut down days (TWC8\_1 to TWC8\_4) - included in CCHS.
- Move: Question on having a regular medical doctor (TWC8\_5) - moved to Health Care Utilisation section.

### **Changes to focus content**

Cycle 3 focus content sections on self-care, family medical history, nutrition, and sense of coherence were dropped. For Cycle 4 the following focus content has been included:

### **CC - Chronic condition (diagnosis and management)**

#### **Arthritis**

- Addition: Question on kind of arthritis (CC\_0\_D1).
- Addition: Question on had operation for arthritis (CC\_0\_D2).
- Addition: Question on mark all types of operation for arthritis (CC\_0\_D2A to CC\_0\_D2D).
- Addition / revised: Question on treatment to relieve arthritis pain (CC\_0\_D3).
- Addition / revised: Questions on mark all types of treatment (CC\_0\_D3A to CC\_0\_D3H).
- Addition: Question on mark all types of drugs (CC\_0\_D4A to CC\_0\_D4C).

#### **Diabetes**

- Addition: Question on pregnant when diabetes diagnosed (CC\_0\_J3A).
- Addition: Question on has non-gestational diabetes (CC\_0\_J3B).
- Addition: Question on insulin started after diagnosis (CC\_0\_J3C).
- Addition: Question on insulin taken daily (CC\_0\_J4).

- Addition: Question on number of times insulin taken each day (CC\_0\_J4A).
- Addition: Question on number of units of insulin per day (CC\_0\_J4B).
- Addition: Question on respondent has taken course on diabetes management (CC\_0\_J5A).
- Addition: Question on someone in household has taken course on diabetes management (CC\_0\_J5B).
- Addition: Question on given information on eating (CC\_0\_J5C).
- Addition: Question on mark all type of health professional giving eating advice (CC\_0\_J6A to CC\_0\_J6E).
- Addition: Question on respondent taught how to check blood sugar levels (CC\_0\_J7A).
- Addition: Question on someone in household was taught how to check blood sugar levels (CC\_0\_J7B).
- Addition: Question on frequency of checking blood sugar levels (CC\_0\_J8A).
- Addition: Question on frequency of checking feet (CC\_0\_J8B).

### **Heart disease**

- Addition: Question on ever had heart attack (CC\_0\_L1A).
- Addition: Question on number of heart attacks (CC\_0\_L1B).
- Addition: Question on age at first heart attack (CC\_0\_L1C).
- Addition: Question on age at most recent heart attack (CC\_0\_L1D).
- Addition: Question on hospitalized for heart attack (CC\_0\_L1E).
- Addition: Question on mark all treatments received for heart attack (CC\_0\_L4A to CC\_0\_L4E).
- Addition: Question on ever referred to cardiac rehabilitation program (CC\_0\_L5A).
- Addition: Question on ever attended cardiac rehabilitation program (CC\_0\_L5B).
- Addition: Question on completed cardiac rehabilitation program (CC\_0\_L5C).
- Addition: Question on has angina (CC\_0\_L6).
- Addition: Question on has congestive heart failure (CC\_0\_L7).

### **ST - Stress**

These questions are a repeat of the questions asked in Cycle 1 with the following changes:

#### **Ongoing problems**

- No change.

#### **Recent life events**

- No change.

#### **Childhood and adult stressors (“traumas”)**

- Revision of universe: Questions asked only of persons who were < 18 in Cycle 3 and now are 18 years and older (data previously collected for those 18 and over in Cycle 1).

**Work stress**

- Revision of universe: Question changed from persons aged 15 and older to persons aged 15 to 75 (now same population used in labour force questions).
- Addition: Question on currently work at a job or business (ST\_0\_W1).

**Self-esteem**

- No change.

**Mastery**

- No change.

**Changes to buy-in content**

There are no buy-in questions in Cycle 4. The following Cycle 3 buy-in questions were dropped:

**FI - Food insecurity**

- Deletion: Question on worry about lack of money for food (FIC8\_1).
- Deletion: Question on not having enough food because of lack of money (FIC8\_2).
- Deletion: Question on not having enough variety or quality of food because of lack of money (FIC8\_3).

**Appendix C: NPHS Household Component, Examples of Cycle 3 (1998-1999) Data Feedback and Follow-up Questions**

<p>Blood Pressure; Mammography; Pap Smear Test</p>	<p>In Cycle 1 and Cycle 2 the respondent was asked whether he/she ever had his or her blood pressure taken (or ever had a mammography etc.). In Cycle 3 the questions were repeated; however, the respondent was probed when said that he or she has not had the test done and in the previous cycle reported the contrary. In Cycle 4, if the respondent had reported that he or she had had the test performed in a previous interview, only the question on the last time it was done was asked.</p>
<p>Restriction of Activities</p>	<p>Information on whether or not the respondent had a disability in Cycle 3 was used in Cycle 4. If the status changed, an explanation of that change was probed.</p>
<p>Chronic Conditions</p>	<p>For all respondents, selected chronic conditions (asthma, 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, the date of onset for the condition was acquired.</p>
<p>Smoking</p>	<p>If a daily smoker had reported the age he or she started smoking daily during last interview, that response was fed back in Cycle 4. For the occasional smoker or non-smoker in Cycle 4 who had reported smoking daily (or having ever smoked daily) during last interview, a flag about daily smoking was re-input. If smoking status changed, an explanation of that change was probed.</p>
<p>Socio-demographic Characteristics</p>	<p>For all respondents, a flag indicating that country of birth had been collected was input again. The ethnic origin and race variables were re-collected. Language first learned and still spoken was asked again because it can change over time.</p>

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 of the main job in Cycle 3 were fed back. If the respondent indicated that they worked in the previous year, they were asked to confirm the employer name.
Health Number	There was a flag that indicated whether the health number that was collected in an earlier interview was valid. If the respondent's health number had not changed since last cycle and was invalid then the health number was asked again.