NATIONAL POPULATION HEALTH SURVEY

HOUSEHOLD COMPONENT

CYCLE 5 (2002-2003)

LONGITUDINAL DOCUMENTATION

STATISTICS CANADA

November 24, 2004

NOTE TO USERS

The National Population Health Survey (NPHS) Cycle 5 longitudinal documentation provides a wide range of information on the survey: objectives, content development, sample design, collection, processing, weighting procedures, data quality, tabulation guidelines and data access. Chapters 7 and 8 give more details on the various subsets of respondents and their associated weights.

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
Cycle 5 = 2002-2003

This guide is also intended for users of the share file, provincial health departments and Health Canada. The share file includes the Cycle 5 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.

Table of Contents

Tables list

Appendix list

Other Reference Documents

Questionnaire
Record Layout
Alphabetic Index
Topical Index
Data Dictionary
NPHS Derived Variables Documentation, Cycles 1 to 5

1. Introduction

The National Population Health Survey (NPHS) is designed to collect "longitudinal" 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 10 cycles. The NPHS fulfilled both cross-sectional and longitudinal needs during its first three cycles, and then with Cycle 4 (2000-2001) the NPHS Household component became strictly a longitudinal survey. The cross-sectional component of the Population Health Surveys Program has been taken over by the Canadian Community Health Survey (CCHS). The NPHS is now composed of two components: the household and the health institutions surveys. Starting with Cycle 4, the survey of the North was strictly cross-sectional and was conducted by CCHS rather than NPHS.

The target population NPHS Household component includes household residents in all provinces in 1994-1995 excluding Indian Reserves and Crown Lands, health institutions, Canadian Forces Bases, and some remote areas in Ontario and Quebec. The Health Institutions component includes long-term residents (expected stay of longer than six months) in health facilities, with four or more beds in all provinces, but excluded the territories, Indian Reserves and Canadian Forces Bases. The Household component of NPHS has completed five cycles: NPHS Cycle 1 (1994-1995), NPHS Cycle 2 (1996-1997), NPHS Cycle 3 (1998-1999), NPHS Cycle 4 (2000-2001) and NPHS Cycle 5 (2002-2003).

The Cycle 5 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 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, household income and labour force status.

This document has been produced to facilitate the use of the Cycle 5 Longitudinal Master and Share Files from the Household component. These files are described in more detail in the following chapters. Any questions about the data sets or their use should be directed to:

Survey content and access to NPHS master files:
Health Statistics Division

Tel: 1-613-951-1653
Fax: 1-613-951-4198
E-mail : nphs-ensp@statcan.ca

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

Tel: 1-613-951-1746
E-mail : hd-ds@statcan.ca

2. Background

In the Fall of 1991, the National Health Information Council (NHIC) recommended that an ongoing national survey of population health be conducted. This recommendation was based on consideration of the economic and fiscal pressures on the health care systems and the commensurate requirement for information 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.

Beginning 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 population's health status;

- 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 or sample;

- allow the possibility of linking survey data to administrative data that are routinely collected, such as vital statistics, environmental measures, community variables, and health services utilization.

4. Survey Content

The above noted objectives 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 the questionnaire sections. The next section describes detailed changes to existing content as well as new content for Cycle 5. The last section discusses Cycle 4 (2000-2001) variables that were fed back and used in Cycle 5. More detailed information on the Cycle 5 content as well as a summary of changes to the content through the first four cycles are available in Appendixes 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 in Cycle 1, 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). From Cycle 4 onwards, the General and Health component questionnaires were combined into a unique questionnaire. Now, the longitudinal respondent answers to the household related questions.

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 is measured through questions on self-perception of health, functional ability, chronic conditions, and activity restriction. The use of health services is 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 are explored include smoking, alcohol use and physical activity. Questions are 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 5 (2002-2003) questionnaire is provided with the documentation and is also available on Statistics Canada's Web site http://www.statcan.ca. Click on "Definitions, data sources and methods", then "Questionnaires, List by subject", then "Health, Questionnaires" and finally "National Population Health Survey, Household Component - Cycle 5 (2002-2003)". Please see Chapter 12 for more details on information available on Statistics Canada's Web site.

4.2 Cycle 5 (2002-2003) Changes to Existing Content

The order of the questions on the questionnaire did not change from cycle 4. Questions on health are first asked and followed by the socio-economic questions (language, education, labour force status, and income).

The Cycle 5 focus and the buy-in contents have been incorporated to the most suitable place of the questionnaire. The cycle 5 focus content was on sleep, body image, nutrition, attitudes toward alcohol and alcohol dependence. The buy-in content was on residence history in the past 20 years. A few questions on the diagnosis and how chronic conditions are managed were reformulated or removed.

For more information, please see Appendixes A and B. Appendix A shows the NPHS first five cycle questionnaire content. Appendix B details changes over 5 cycles in the NPHS questionnaire.

4.3 Cycle 4 (2000-2001) 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 repeated. Variables that could change over time if certain actions had occurred (e.g., level of education), were updated was only if appropriate. Some answers from earlier cycles were brought forward into the Cycle 5 interview. This proved to be a valuable tool resulting in better quality collected data. For instance, previous information on selected chronic conditions was recalled for the respondent in order to explain any change between Cycle 4 and Cycle 5 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 institutions, full-time members of the Canadian Forces Bases and some remote areas in Ontario and Quebec. This chapter describes the Cycle 1 sample design and explains how the sample of 17,276 persons was selected.

5.1 Cycle 1 (1994-1995) Sample Design

The Labour Force Survey (LFS) sample design, redesigned in 1991, was used as the basis for the sample design in all provinces except Quebec where the NPHS sample was selected 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 LFS sample design as a vehicle for selecting the sample.

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

5.1.1 Sample Allocation

The NPHS initially had a target sample size of 19,600 households. It was further agreed by 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)[1]. According to this scheme the sample was allocated proportionally to $\sqrt{(0.804 W_h{}^2 + 1/12^2)}$, where W_h is the 1991 Census proportion of households in province/territory (h, h=1,..,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 option of buy-in of additional units with increased funding, for cross-sectional purposes. These additional units were not retained for the longitudinal sample.

---

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

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 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. To do so, 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 preliminary 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 a dwelling list was prepared for each chosen cluster, and some were selected from the list.

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 this was 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 representativeness (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 (NLSCY), 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. It also includes 2,022 children from the first cycle (1994–1995) of the National Longitudinal Survey of Children and Youth (NLSCY).

These children were interviewed by the NLSCY for the NPHS in Cycle 1 and are interviewed by the NPHS since the second cycle. This panel, surveyed in Cycles 2, 3, 4 and 5, 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 5, all longitudinal panel members were 8 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 9.2.4), 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 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 provided a full response to all five cycles of NPHS (this count includes the deceased persons).

Table 5.A: Longitudinal Sample Size by Province

Province	Longitudinal Sample Cycle 1 (1994-1995)	Records Providing a Full Response in Cycles 1, 2, 3, 4 and 5
Newfoundland	1,082	822
Prince Edward Island	1,037	803
Nova Scotia	1,085	775
New Brunswick	1,125	824
Quebec	3,000	2,189
Ontario	4,307	2,990
Manitoba	1,205	921
Saskatchewan	1,168	922
Alberta	1,544	1,111
British Columbia	1,723	1,189
Total	**17,276**	**12,546**

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.

6.3   Interviewing

In Cycle 5, collection for the household sample was divided into four quarters (starting in June, August and October 2002, and January 2003). An additional collection period was held in June 2003 with further follow-up of non-respondents from previous quarters.

The interviewers were 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 5 data collection. They also received a brochure that presents some results extracted from the survey data and a magnet bearing the Population Health Surveys logo to thank respondents for answering the questions.

In general, respondents in the household sample are contacted by telephone, and over 98 % of the interviews in Cycle 5 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 just under an hour.

Information about all household members (age, sex, and relationships between members) was obtained from the longitudinal respondent. Proxy reporting for the longitudinal respondent aged 12 and over was allowed only for reasons of illness or incapacity. Such proxy reporting accounted for 4.8 % 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

Many strategies were put in place to reduce the number of non-response cases. For example, 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.

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 respondent's co-operation. This was followed by a second call (or visit) from the interviewer. If the time of the call (or visit) was not adequate, the interviewer agreed with the respondent on the best time to call back (or to come back).

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 other contacts, if collected in a previous cycle. In addition, interviewers were trained to follow up available public leads such as local telephone directories and directory assistance. If these leads were unsuccessful, the case was transmitted to an interviewer specially trained in tracing respondents. Tracer interviewers had access to Canada-wide telephone directories and reverse directories. The cumulative non-response rate due to failure to trace the longitudinal respondent is 4.1 % of the total panel, which is relatively low for the fifth cycle of the survey. Section 9.2.5 presents non-response rates due to non-tracing with more details.

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 outside the United States; information was updated for next cycle.

7. Data Processing

7.1 Editing

Editing was first 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 the contradiction was accepted and 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 Residence History, a new module in Cycle 5

A separate file for the Residence History microdata has been created. A specific record layout and data dictionary are available for this separate file. During data collection, respondents had the option of reporting residential moves either forward from 1980 or backwards to 1980. During processing all data collected was converted to one direction only (backwards to 1980).

7.3 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.

In Cycle 5 master data file (which contains all 17,276 longitudinal panel members), the industry and occupation data for all cycles are coded to the North American Industrial Classification System (NAICS) and Standard Occupational Classification 1991 (SOC-91).

The drug coding is based on the Anatomical Therapeutic Chemical (ATC) classification developed by the World Health Organisation (WHO) as available on the Health Canada Drug Product Database (DPD) in September 2003. For Cycle 5, a complete revision of the drug codes was done for all NPHS cycles. A complete list of the codes is available upon request.

Conditions or health problems causing activity restrictions were coded based on the International Classification of Diseases, 9th Revision (ICD-9) or according to the Musculoskeletal Impairment Supplementary Coding Scheme developed for the Health and Activity Limitation Survey (HALS).

7.4 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 11.3 for more detail on the variable naming conventions). 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 5" 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. In some cases, the derived variables are straightforward, involving collapsing response categories. Grouped variable names generally have a "G" in the fifth character of the variable name (see Section 11.3 for more detail on the variable naming conventions).

7.5 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 8.

7.6 Subsets of respondents

In order to provide greater flexibility to users, a single microdata master file is being issued for NPHS Cycle 5. This file includes all 17,276 NPHS panel members, notwithstanding their response patterns from Cycles 1 to 5. 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 8 for more information regarding the calculation of each subset's sampling weights and to Section 11.1 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

Subset of Respondents	Type of Response	Flags	Number of Respondents
Longitudinal Square	Complete panel: all panel members regardless of their response pattern.	None, all records	17,276
Longitudinal Full	All panel members with a complete response (Full) in Cycles 1, 2, 3, 4 and 5.	WF62LF	12,546
Longitudinal Full C1 and C5	All panel members with a complete response (Full) in Cycles 1 and 5 regardless of their response pattern in Cycles 2, 3 and 4.	WF62LFE	13,629
Longitudinal Full Share	All panel members with a complete response (Full) in Cycles 1, 2, 3, 4 and 5 and who agreed to share their data in Cycle 5.	WF62SLF	12,226

Users of the share file, provincial health departments and Health Canada, 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.

7.7 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 and the date of death (if not collected during the survey) are 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 Institutions survey. Their data were brought back to the NPHS household component dataset. However, the health institutions component collects less information then the household component. The missing variables are set to "6" (i.e., not applicable).

8. Weighting

This chapter describes the weighting procedures for each subset of respondents described in Section 7.6. 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, 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. 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. Although panel members were part of the cross-sectional and longitudinal files, their weights were not identical for these two types of files but rather adjusted to correctly represent the target population.

For Cycle 5, four sets of weights, WT64LS, WT62LF, WT62LFE and WT62SLF have been constructed. Table 8.A shows the subsets of respondents and the corresponding sampling weights and flags. A panel member is part of a given subset when the flag is equal to 1.

Table 8.A: Subsets of Respondents and Corresponding Sampling Weights and Flags

Subsets of respondents	Sampling Weights	Flags
Longitudinal Square	W**T**64LS	None, all records
Longitudinal Full	W**T**62LF	W**F**62LF
Longitudinal Full C1 and C5	W**T**62LFE	W**F**62LFE
Longitudinal Full Share	W**T**62SLF	W**F**62SLF

The four weights were post stratified to the 1994-1995 population estimates based on the 1996 Census counts by age group and sex within each province. The WT62LF, WT62LFE and WT62SLF weights have been adjusted for non-response. 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.

8.1 Longitudinal Weighting

The longitudinal weighting procedure is based on the weighting done for the Cycle 1 NPHS cross-sectional sample. Some weight adjustments were applied to the Cycle 1 cross-sectional weights in order to incorporate the additional sample used exclusively for cross-sectional purposes. These adjustments were removed for the longitudinal panel weight to create a "stripped" weight. This stripped weight is the starting point to obtain the longitudinal weight.

8.1.1 Starting point: 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.

8.1.2 Adjustments to create the different square weights

8.1.2.1 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:

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

8.1.2.2 Longitudinal Full Weight (WT62LF)

The longitudinal full subset includes only selected members who responded in all cycles or who have died. Panel members that are excluded from this subset were therefore non-respondents at some point during the first five cycles of the survey, and their weight must be redistributed to compensate for this 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 5 weights, the non-response adjustments for Cycles 2, 3, 4 and 5 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 8.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 8.B.

Table 8.B: Variables for Cycle 2 Non-response adjustment

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 within each weighting class.

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 (LONGPAT variable) 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=non-response. 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 8.C. A 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 8.C.

Table 8.C: Variables for Cycle 3 / Non-response Adjustment

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^E
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:

  Sum of weights for Cycles 1 and 2 responding longitudinal members
 Sum of weights for Cycles 1, 2 and 3 responding longitudinal members

This adjustment is performed within each weighting class, 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 8.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 8.D.

Table 8.D: Variables for Cycle 4 / Non-response Adjustment

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:

Sum of weights for Cycles 1, 2 and 3 responding longitudinal members
  Sum of weights for Cycles 1 to 4 responding longitudinal members

This adjustment is performed within each weighting class. 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: Adjustment for Cycle 5 (2002-2003) Non-response

The 13,582 records with a full response after four cycles are taken as the starting point. Once again, records for which the panel member was deceased in Cycle 2, 3 or 4 or institutionalized since Cycle 2, 3 or 4 are treated differently from the rest. For these records, no non-response adjustment is made since their weight in Cycle 2, 3 or 4 has been already adjusted to reflect the fact that some of the Cycle 2, Cycle 3 or Cycle 4 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 5, only the Cycle 4 characteristics of the household as well as personal characteristics of the longitudinal member were considered. As for Cycle 4, 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 5 non-response adjustment.

The variables chosen by the CHAID algorithm to build the weighting classes to adjust for Cycle 5 non-response are in Table 8.E. Three Cycle 1 design variables were also used, identifying the presence of household members under the age of 12, under the age of 25, and the "adult/child" household type. A Cycle 4 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 8.E.

Table 8.E: Variables for Cycle 5 / Non-response Adjustment

ALC0_3	DHC0_OWN	IMM	MHC0_1J	SMC0_2
ALC0DTYP	DHC0DL12	INC0DIA5	MHC0DCH	ST_0DC4
ALC0DWKY	DHC0DLE5	ISC0_1	MHC0DDS	ST_0DC5
AM60_SHA	GE30DURB	LSC0_1	PAC0DFD	ST_0DC6
BPC0_10	GHC0_21	LSC0DPFT	PAC0DLEI	ST_0DC8
CCC0DANY	HCC0DHPC	MHC0_16	SDC0_4A	ST_0DR2
DGC0F1	HSC0DHSI	MHC0_1A	SDC0_6A	ST_0DW3
DHC0_AGE	HWC0DSW	MHC0_1F	SEX	ST_0DW6

To adjust for longitudinal members that did not respond in Cycle 5, the following adjustment is applied to the weight of respondents:

Sum of weights for Cycles 1, 2, 3 and 4 responding longitudinal members
Sum of weights for Cycles 1 to 5 responding longitudinal members

This adjustment is performed at the weighting class level. Records for which the panel member was deceased in Cycle 2, 3 or 4 or institutionalized since Cycle 2, 3 or 4 are not part of this adjustment.

Adjustment 5: Post-stratification adjustment

The weight of the units that are part of the subset were 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

<u>Population estimate in a province/age/sex category</u>
Sum of weights of Cycles 1 to 5 responding longitudinal members in a province/age/sex category

The final longitudinal weight **WT62LF** is calculated by taking the Cycle 1 stripped weight and multiplying that value by Adjustments 1 to 5.

8.1.2.3 Longitudinal Full Weight for Cycles 1 and 5 (WT62LFE)

The NPHS now includes, starting with Cycle 4, an additional weight to be used with the subset of panel members who responded to both Cycle 1 and the most recent cycle (in this case, Cycle 5). This subset includes the respondents that are part of the full subset, as well as those respondents with a full response in Cycles 1 and 5 only (that is, some non-respondents and partial respondents in Cycles 2, 3 or 4). 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 5 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 non-respondents in Cycles 2, 3 or 4 whose weights had been distributed to the respondents that make up the full subset were respondents once again in Cycle 5 and their weights now have to be retrieved. The starting point for each retrieved record in Cycle 5 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 5 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 5 represent the same proportion in the full response in Cycles 1 and 5 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 full response in Cycles 1 and 5 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 :

$$\frac{\text{Population estimate in a province/age/sex category}}{\substack{\text{Sum of weights of the Full C1 and C5 subset members} \\ \text{in a province/age/sex category}}}$$

The final weight for a full response in Cycles 1 and 5 is the weight WT62LFE. For a more technical and detailed description of the adjustments used in the creation of this weight, please refer to Brisebois and Mathieu (2003) [2].

8.1.2.4 Longitudinal Full Share Weight (WT62SLF)

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:

$$\frac{\substack{\text{Sum of weights for Cycles 1 to 5 responding longitudinal members} \\ \text{in a province / longitudinal pattern / age-sex category}}}{\substack{\text{Sum of weights for Cycles 1 to 5 responding longitudinal members who} \\ \text{agreed to share, in a province / longitudinal pattern / age-sex category}}}$$

Note that in Cycles 3, 4 and 5; 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

---

[2] Brisebois, F. and Mathieu, P. (2003) Creation of a new longitudinal weight for the Canadian National Population Health Survey: Providing data users with greater analytical flexibility. Proceedings of the Survey Methods Section, Statistical Society of Canada

deceased for each of these cycles. The final longitudinal sharing weight (WT62SLF) is obtained by multiplying the longitudinal full weight (WT62LF) by this adjustment. Note that since this adjustment is done with respect to the post-stratification classes, no additional post-stratification is necessary.

9. Data Quality

Data quality is an important aspect for any survey. Examining data quality allows the verification of the reliability and accuracy of the data collected, as well as help to determine what steps should be taken to improve data quality in future cycles.

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 misunderstand the questions asked, the answers may be incorrectly entered or errors may be introduced in the processing and tabulation of the data. These are all examples of *non-sampling errors*.

9.1   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:

$$(.003/.25) \text{ x } 100\% = 1.20\%.$$

Statistics Canada commonly uses CV results to verify the quality of statistical estimates produced when analyzing data, and strongly 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 10.4.

9.1.1   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 as well as a large number of statistical tests 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 approximate 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, 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. For more information on Bootstrap weights, please refer to Section 11.2.

The Bootvar program is available in both SAS and SPSS formats. It is made up of macros 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.

9.2   Non-sampling Errors

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 is minimal; once the questionnaire is started, it tends 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 or reach 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 8.1.2 for details of the weight adjustment for non-response.

This section presents some information dealing with different aspects of non-response. Discussed first is overall non-response, where non-response rates from each cycle are presented. This is followed by sections related to refusals, refusal conversion, attrition, and the tracing of longitudinal respondents. Finally, item non-response is briefly examined.

9.2.1 Response 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. 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 Cycles 4 and 5.

9.2.1.1 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)[3] were excluded from the panel and from the calculations of the Cycle 1 response rates.

Selected-person response rate for H05

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

---

[3] 017 = Other ineligible dwelling (e.g., embassy).
  018 = Rejected household.
  023 = Under construction or demolished.
  024 = Vacant dwelling.

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.

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:

$$\text{Selected-person H05 Rate} = \frac{17,276}{17,276 + 2,819} = \frac{17,276}{20,095} = 86.0\%$$

$$\text{Selected-person H06 Rate} = \frac{16,794}{16,794 + 3,301} = \frac{16,794}{20,095} = 83.6\%$$

9.2.1.2 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 who have moved to and been interviewed in a health institution are counted as a response for longitudinal purposes; no panel members are classified as out-of-scope.

Panel response rate for H05

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

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

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

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[4]:	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:

Panel response rate to the H05 component =

$$\frac{15{,}819 + 287 + 62}{15{,}819 + 287 + 62 + 1{,}108} = \frac{16{,}168}{17{,}276} = 93.6\%$$

Panel response rate to the H06 component =

$$\frac{15{,}687 + 287 + 62}{15{,}687 + 287 + 62 + 1{,}240} = \frac{16{,}036}{17{,}276} = 92.8\%$$

9.2.1.3 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 who have moved to and been interviewed in a health institution are counted as a response for longitudinal purposes; no panel members are classified as out-of-scope.

---

[4] 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.

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.

Relevant information for calculation of response rates:

Number of longitudinal panel members:	17,276
Number of panel members who have died[5]:	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:

Panel response rate to the H05 component =

$$\frac{14,647 + 599 + 114}{14,647 + 599 + 114 + 1,916} = \frac{15,360}{17,276} = 88.9\%$$

Panel response rate to the H06 component =

$$\frac{14,532 + 599 + 114}{14,532 + 599 + 114 + 2,031} = \frac{15,245}{17,276} = 88.2\%$$

---

[5] 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.

9.2.1.4 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 who have moved to and been interviewed in a health institution 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

# of panel members responding or who have died or been institutionalized
                         # 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.

Relevant information for calculation of response rates:

Number of longitudinal panel members:	17,276
Number of panel members who have died[6] :	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:

Response rate for panel=

$$\frac{13,559 + 957 + 135}{13,559 + 957 + 135 + 2,625} \quad = \quad \frac{14,651}{17,276} \quad = 84.8\%$$

9.2.1.5 Cycle 5 (2002-2003) Response Rates

As for Cycles 2, 3 and 4, the Cycle 5 longitudinal response rate is based on the 17,276 members of the longitudinal panel. Persons who have died or who have moved to and been interviewed in a health institution are counted as a response for longitudinal purposes; no panel members are classified as out-of-scope. As NPHS is now purely longitudinal there is no longer

---

6 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 5.

distinction between the H05 questionnaire and 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 ***80.6%*** at the Canada level. At the provincial level, this rate varied from 77.7% in (British Columbia) to 84.7% in (Saskatchewan).

Relevant information for calculation of response rates:

Number of longitudinal panel members:	17,276
Number of panel members who have died:	1,279
Number of panel members who have been institutionalized:	161
Number of respondent panel members:	12,484
Number of non-respondent panel members:	3,352

Calculation of the panel response rate:

Response rate for panel=
$$\frac{12{,}484 + 1{,}279 + 161}{12{,}484 + 1{,}279 + 161 + 3{,}352} = \frac{13{,}924}{17{,}276} = 80.6\%$$

9.2.2 Refusal Rates

Refusals are the most substantial source of nonresponse for the NPHS. Refusals make up 49% of Cycle 2 nonresponse, 56% of Cycle 3 nonresponse, and 61% of the nonresponse in Cycles 4 and 5. Even though the intention is to follow all 17,276 panel members over time, not all records are sent out for collection each cycle, such as the more difficult refusals. Note that cases where the panel member has been confirmed dead through a match to the mortality files are considered complete for the rest of the span of the survey, and are no longer sent out for collection.

Two different refusal rates for each cycle can be calculated, one based only on those records that were sent out, and the other based on all 17,276 records. It can be seen in Table 9.A, which displays both of these rates for Cycles 2, 3, 4 and 5, that both refusal rates increased with each cycle.

Table 9.A – Refusal Rates by Cycle

	Number of records that went out	Number of new Refusals	Refusal rate based on records sent out	Number of refusals that were not sent out	Total number of refusals	Refusal rate based on all 17,276
Cycle 2	17,266	539	3.1%	1	540	3.1%
Cycle 3	16,582	601	3.6%	469	1070	6.2%
Cycle 4	16,186	1017	6.3%	526	1543	8.9%
Cycle 5	15,616	1265	8.1%	687	1952	11.3%

9.2.3 Refusal Conversion Rates

Table 9.B shows the conversion rates between cycles. The refusal conversion rate between cycles is the percentage of respondents that refused in a particular cycle and provided a full response in a subsequent cycle. These rates are based only on the records that were sent out for collection, since only those records that were sent out could be converted, therefore the number of refusals in Table 9.B differs from the number of refusals in Table 9.A. For example, the refusal conversion rate from Cycle 2 to Cycle 3 is based on only 73 records, as 467 of the 540 refusals in Cycle 2 were not sent out for collection in Cycle 3.

Table 9.B: Refusal Conversion Rates

	Number of Refusals	Converted Next Cycle	Converted 2 Cycles Later	Converted 3 Cycles Later	Not Converted
Cycle 2	73	22 (30%)	16 (22%)	2 (3%)	33 (45%)
Cycle 3	539	164 (30%)	34 (6%)		341 (63%)
Refusals to Cycles 2 & 3	26	8 (31%)	0 (0%)		18 (69%)
Cycle 4	849	164 (19%)			685 (81%)
Refusals to Cycles 3 & 4	259	26 (10%)			233 (90%)
Refusals to Cycles 2, 3 & 4	15	0 (0%)			15 (100%)

It can be seen in Table 9.B that conversion rates are much higher over the short term than the long term.

9.2.4 Attrition Rates

In a longitudinal survey, attrition is a loss in sample size due to non-respondents, movements out-of-scope and untraceable individuals. In this section, attrition is defined by whether or not a panel member is part of the full subset. Therefore, when a nonresponse is observed for a panel member, it is part of attrition. Two different attrition rates are being calculated: one showing the attrition rate between two consecutive cycles, 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 information concerning the full subset, see Section 7.7).

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	15,666
Number of individuals in the Cycle 3 (1998-1999) Full subset	14,618
Number of individuals in the Cycle 4 (2000-2001) Full subset	13,582
Number of individuals in the Cycle 5 (2002-2003) Full subset	12,546

Attrition rates between two cycles

$$\text{Cycle 2 (1996-1997):} \quad \frac{17{,}276-15{,}666}{17{,}276} = \frac{1610}{17{,}276} = 9.3\%$$

$$\text{Cycle 3 (1998-1999):} \quad \frac{15{,}666-14{,}618}{15{,}666} = \frac{1048}{15{,}666} = 6.7\%$$

$$\text{Cycle 4 (2000-2001):} \quad \frac{14{,}618-13{,}582}{14{,}618} = \frac{1036}{14{,}618} = 7.1\%$$

$$\text{Cycle 5 (2002-2003):} \quad \frac{13{,}582-12{,}546}{13{,}582} = \frac{1036}{13{,}582} = 7.6\%$$

Cumulative attrition rates

Cycle 2 (1996-1997): $\dfrac{17,276-15,666}{17,276}$ = $\dfrac{1610}{17,276}$ =   9.3%

Cycle 3 (1998-1999): $\dfrac{17,276-14,618}{17,276}$ = $\dfrac{2658}{17,276}$ =   15.4%

Cycle 4 (2000-2001): $\dfrac{17,276-13,582}{17,276}$ = $\dfrac{3694}{17,276}$ =   21.4%

Cycle 5 (2002-2003): $\dfrac{17,276-12,546}{17,276}$ = $\dfrac{4730}{17,276}$ =   27.4%

As is typically the case in longitudinal surveys, the attrition rate between Cycles 1 and 2 is considerably higher (9.3%) than those subsequently observed. The subsequent attrition rates are more constant between cycles. Cumulatively, more than one quarter of the panel is part of the attrition after five cycles.

Table 9.C: Attrition Type by cycle - Full subset of respondents

Attrition type	Cycle 2		Cycle 3		Cycle 4		Cycle 4		Total	
	Numbers	%	Numbers	%	Numbers	%	Numbers	%	Numbers	%
Partial data C1	482	29.9	s/o	s/o	s/o	s/o	s/o	s/o	482	10.2
Partial data	110	6.8	93	8.8	61	5.9	86	8.3	350	7.4
Moved	75	4.7	55	5.2	53	5.1	20	1.9	203	4.3
Refusal	492	30.6	474	45.2	541	52.2	523	50.5	2030	42.9
Unable to trace	275	17.1	131	12.5	175	16.9	162	15.6	743	15.7
Other nonresponse	176	10.9	295	28.1	206	19.9	245	23.6	922	19.5
Total	**1610**		**1048**		**1036**		**1036**		**4730**	**100**

9.2.5 Non-response 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 9.D presents the tracing profile of the panel members according to whether they were traced or not in Cycles 2, 3, 4 and 5.

Table 9.D: Tracing for Cycles 2, 3, 4 and 5

Cycle 2	Cycle 3	Cycle 4	Cycle 5	Number of panel members
T	T	T	T	16,230
T	T	T	UT	325
T	T	UT	T	120
T	T	UT	UT	136
T	UT	T	T	56
T	UT	T	UT	17
T	UT	UT	T	24
T	UT	UT	UT	72
UT	T	T	T	66
UT	T	T	UT	18
UT	T	UT	T	6
UT	T	UT	UT	13
UT	UT	T	T	43
UT	UT	T	UT	17
UT	UT	UT	T	24
UT	UT	UT	UT	109
Total				17,276

T: Traced
UT: Untraced

The longer a panel member stays untraced, the less likely it is that they will be traced in the next cycle.

Using the data from the survey, it was also observed that non-respondents in one cycle are four times more likely than respondents to be untraced. Around 3.5% of the cycle 2 non-respondents were unable to be traced in Cycle 3, 4.6% of the Cycle 3 non-respondents were untraced in Cycle 4, and 7.2% of the Cycle 4 non-respondents were untraced in Cycle 5.

Table 9.D shows that tracing efforts are very valuable as even some panel members that were untraceable for three consecutive cycles were traced in Cycle 5.

9.2.6 Item Refusal and Don't Know Rates

Refusal rates by variable vary between 0% and 5%. Module refusal rates vary between 0% and 0.9%. It tends to be the same variables and modules that have relatively high refusal rates in each cycle. For example, the income variables have refusal rates of close to 4%. Some of the labour force submodules and the mastery

submodule (stress module) have the highest refusal rates at 0.11% and 0.12% respectively.

Don't Know rates by variable vary between 0.0 % and 72.7 %. Module Don't Know rates vary between 0.01% and 2,74%. It tends to be the same variables and modules that have relatively high Don't Know rates in each cycle. For example, the insurance module and the income module have their highest Don't Know rates at 2.0% and 1.8% respectively.

Refusal and Don't know rates were also calculated at the respondent level to determine the percentage of questions an individual refuses to answer or answers Don't Know. Table 9.E shows the proportion of panel members who did not refuse any questions, who refused less than 1% of the questions asked, and who refused less than 3% of the questions asked.

Table 9.E – Refusal and Don't Know Rates at the Respondent Level

	Refusals to 0% of questions	Refusals to less than 1% of questions	Refusals to less than 3% of questions	Don't know to 0% of questions	Don't know to less than 1% of questions	Don't know to less than 5% of questions
Overall	92.3%	98.1%	98.8%	62.6%	93.7%	99.4%
Males	92.2%	97.7%	98.6%	63.7%	93.9%	99.3%
Females	92.3%	98.5%	98.9%	61.6%	93.6%	99.5%
Under 12	94.5%	95.7%	98.0%	82.9%	95.5%	99.5%
12-24	93.2%	98.5%	98.6%	46.5%	88.6%	98.3%
25-44	94.4%	98.9%	99.2%	72.4%	97.6%	99.8%
45-64	92.4%	98.5%	99.0%	65.9%	96.3%	99.7%
65+	86.8%	96.6%	97.9%	48.0%	86.6%	99.0%
Proxy	91.6%	95.9%	97.9%	69.5%	88.3%	98.6%
Non-Proxy	92.3%	98.3%	98.8%	61.9%	94.2%	99.5%

It can also be seen that don't know rates have more variation than the refusal rates when they are separated by sex, age group and interview type. Of the respondents who fully completed the interview in Cycle 5, 92.3% answered all of the questions. Almost 99% have refusal rates less than 3%. As for Don't know rates, 62.6% have Don't know rates of 0%. Over 99% have Don't know rates less than 5%, and almost 94% have Don't know rates less than 1%. This shows that almost everyone who refuses or responds Don't know does so for only a few questions.

10. 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.

10.1 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. Un-rounded 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 un-rounded 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, they are to be rounded 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 un-rounded 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).

10.2 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.

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

10.2.1   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.

Example of Categorical Question:

> *At the present do/does ... smoke cigarettes daily, occasionally or not at all? (**SMC2_2**)*
>   __ Daily
>   __ Occasionally
>   __ Not at all

Quantitative Estimates:

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

An example of a quantitative estimate is the average increase in the number of cigarettes smoked per day by daily smokers who had an increase in consumption between two cycles.

Example of Quantitative Question:

How many cigarettes do/does you/he/she smoke each day now? (SMC2_4)

|_|_| Number of cigarettes

10.2.2   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)   by summing the final weights of records having the characteristic of interest for the numerator ($\hat{X}$),

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

c)   by dividing the numerator estimate by the denominator estimate.

10.2.3   Tabulation of Quantitative Estimates

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

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

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

c)   dividing the numerator estimate by the denominator estimate.

For example, to obtain the estimate of the average number of cigarettes smoked each day by individuals who smoke daily, first compute the numerator ($\hat{X}$) by summing the product between the value of variable **SMC2_4** and the final weight. Next, sum this value over those records with a value of "daily" to the variable **SMC2_2**. The denominator ($\hat{Y}$) is obtained by summing the final weight of those records with a value of "daily" to the variable **SMC2_2**. Divide ($\hat{X}$) by ($\hat{Y}$) to obtain the average number of cigarettes smoked each day by daily smokers.

10.3 Guidelines for Statistical Analysis

The 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 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.

10.4 Release Guidelines

Before releasing or publishing any total or proportion estimates from the master files, users must first determine the number of sampled respondents having the characteristic of interest (for example, the number of respondents who smoke when interested in the proportion of smokers for a given population). If this number is less than 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 10.A.

Table 10.A: Sampling Variability Guideline

Type of Estimate	C.V. (in %)	Guidelines
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."

11. Using the Longitudinal Master Files

11.1 Use of Longitudinal Weights

The Cycle 5 master file contains 17,276 panel members and four subsets of respondents (Section 7.6) to which correspond a set of weights (Section 8.1). This set of data includes respondents who become non-respondents. Flags were created to identify records that are part of a particular subset (Table 8.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 17,276 members that make up the original longitudinal sample.   All non-response should be taken into account for any calculation.

Weight WT62LF is called the "Longitudinal Full" weight and applies to the 12,546 records that are included in the "Full" subset of respondents.

Weight WT62LFE is called the "Longitudinal Full C1 and C5" weight and applies to the 13,629 records that are included in the "Full C1 and C5" subset of respondents.

Weight WT62SLF is called "Longitudinal Full Share" weight and applies to the 12,226 respondents that are included in the "Full Share" subset of respondents.

11.2 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 9.1.1. Each subset of respondents has a set of bootstrap weights associated with it. Four different sets of bootstrap weights were created for the Cycle 5 data: the square, the full, the full C1 and C5 and the full share. For more information on these subsets, see Section 7.5. Table 11.A presents the subset of respondents with their corresponding bootstrap file name.

Table 11.A: Subsets of Respondents and Corresponding Bootstrap Weights files

Subset of respondents	Name of the Bootstrap Weights file
Longitudinal Square	B5long
Longitudinal Full	B5lngf
Longitudinal Full C1 and C5	B5lngfe
Longitudinal Full Share	B5lngf (share)

Due to the complex sample design, users should use the supplied 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 10 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.

11.3 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, 2000-2001, 2002-2003) 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, 1998-1999, 2000-2001 and in 2002-2003, 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 institutions component, and previously the North component and supplements.

11.3.1 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, DHC0_AGE and DHC2_AGE:

DH:	in the Demographic and Household content section of the questionnaire;
C:	questions which are Core content on the household survey;
4/6/8/0/2:	appeared in Cycle 1 (1994-1995), / appeared in Cycle 2 (1996-1997), / appeared in Cycle 3 (1998-1999), / appeared in Cycle 4 (2000-2001), appeared in Cycle 5 (2002-2003);
_:	can be found on the questionnaire;
AGE:	the variable name.

11.3.2 Positions 1-2: Variable Name / Questionnaire Section Name

AD	Alcohol dependence	MH	Mental health
AL	Alcohol	NU	Nutrition
AM	Administration of the survey	PA	Physical activities
CC	Chronic conditions	PH	Preventive health
DG	Drug use	PY	Psychological resources (self-esteem, mastery, sense of coherence)
DH	Demographics and household	RA	Restriction of activities
ED	Education	RH, MB	Residence history
FV	Fruit and vegetable consumption	RP	Repetitive strain
GE	Geographic identifiers	SD	Socio-demographics
GH	General health	SL	Sleep
HC	Health care utilization	SM	Smoking
HS	Health status	SP	Sample control variables
HW	Height and weight	SS	Social support
IJ	Injuries	ST	Stress
IN	Income	TU	Tanning and UV exposure
IS	Insurance	WF	Subset flags
LF/LS	Labour force	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 11.B presents the variables that appear only once of the data file. The name of these variables does not follow the naming convention.

Table 11.B: "Constant" Longitudinal Variables

Longitudinal Name	Concept
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

11.3.3   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, 4 and 5, 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) (for example: weights, agreement to share, date of interview variables, etc.)

11.3.4 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)

11.3.5 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

11.3.6 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_2_C1, the first question on work stress was named ST_2_W1. 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.

12. Access to NPHS Data

12.1 Microdata Files

Confidentiality concerns preclude general dissemination of longitudinal NPHS data in public use microdata file (PUMF) format. However, access to all the longitudinal master microdata files including the Cycle 4 and Cycle 5 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 Statistics Canada's Research Data Centres (RDCs) program. The RDCs program is part of an initiative by Statistics Canada, the Social Sciences and Humanities Research Council (SSHRC) and university consortia to help strengthen Canada's social research capacity and to support the policy research community.

RDCs provide researchers with access, in a secure university setting, to microdata from population and household surveys. The centres are staffed by Statistics Canada employees. They are operated under the provisions of the *Statistics Act* in accordance with all the confidentiality rules and are accessible only to researchers with approved projects who have been sworn in as "deemed employees". RDCs are located throughout the country, so researchers do not need to travel to Ottawa to access Statistics Canada microdata. More information is available at the Research Data Centre Program web site: http://www.statcan.ca/english/rdc/index.htm.

A second option, if the RDCs are not accessible for the researcher, is 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.

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 or one of Statistics Canada's Regional Offices.

12.2 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, and select Products and Services.

Longitudinal Cansim tables are also available free of charge on the Statistics Canada Internet site. They present changes, from one NPHS cycle to another one, in smoking, self-rated health, body mass index and physical activity. To find the Cansim tables, click on "*Healthy today, healthy tomorrow? Findings from the National Population Health Survey*" (82-618-MWE), then "Data tables".

Custom tabulations from the NPHS are also 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.

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, Questionnaire Content over 5 cycles

Theme	Subject	Variable	Universe	1994-1995 (Cycle 1)	1996-1997 (Cycle 2)	1998-1999 (Cycle 3)	2000-2001 (Cycle 4)	2002-2003 (Cycle 5)
Administration		AM	All respondents	C	C	C	C	C
Alcohol	Attitudes towards	AL	Cycle 2: >=12 Alberta RDD excluded. Cycle 5: >=12		HPS			F
	Consumption during pregnancy	AL	Cycle 1: Females and birth mother in last 5 years. Cycle 2: 15 to 49 years old Alberta only	HPS	PBI			
	Dependence	AD	>=12		F			F
	Use	AL	>=12	C	C	C	C	C
Attitudes	Towards parents	AP	Ages 12 to 17 in Alberta		PBI			
Body image		HW	>=12	HPS	HPS			F
Breast-feeding		BF GH	Cycle 1: Birth mother in the last five year. Cycle 2: HPS: Females >=15 and <=49 and have given birth since last interview; PBI: Alberta	HPS	HPS PBI			
Child's Health		IJ DG	<12		C			
Chronic conditions	Acne requiring prescription medication	CC	<30	C				
	Alzheimer's disease or other dementia	CC	>=18	C	C	C	C	C
	Arthritis or rheumatism	CC	>=12	C	C	C	C F	C
	Asthma	CC	Cycle 1: >=12. Cycle 2-5: All respondents. Cycle 2 - AS: All respondents >=12 who declared having asthma diagnosed by a health professional in the core survey.	C	C AS	C	C	C

C: Core, **F**: Focus, **HPS**: Health Promotion Survey, **AS**: Asthma Supplement, **FIS**: Food Insecurity Supplement, **PBI**: Provincial Buy-in, **K**: Kids' questions, **OBI**: Other Buy-ins.

Theme	Subject	Variable	Universe	1994-1995 (Cycle 1)	1996-1997 (Cycle 2)	1998-1999 (Cycle 3)	2000-2001 (Cycle 4)	2002-2003 (Cycle 5)
Chronic conditions	Back problem	CC	>=12	C	C	C	C	C
	Bowel disorder such as Crohn's disease or colitis	CC	>=12		C	C	C	C
	Bronchitis, emphysema	CC	Cycle 1 and 5: >=12, Cycle 2-4: all respondents	C	C	C	C	C
	Cancer	CC	>=12	C	C	C	C	C
	Cataracts	CC	>=18	C	C	C	C	C
	Child	CC	<=3	K	K	K		
	Diabetes	CC	Cycle 1-4: >=12, Cycle 5: all respondents	C	C	C	C F	C
	Effects of stroke	CC	>=12	C	C	C	C	C
	Epilepsy	CC	Cycle 1: >=12, Cycle 2-5: all respondents	C	C	C	C	C
	Fibromyalgia	CC	>=12				C	C
	Food allergies	CC	Cycle 1: >=12. Cycle 2-5: All respondents	C	C	C	C	C
	Glaucoma	CC	>=18	C	C	C	C	C
	Heart Disease	CC	Cycle 1: >=12, Cycle 2-5: all respondents	C	C	C	C F	C
	High blood pressure	CC	>=12	C	C	C	C	C
	Migraine, headache	CC	Cycle 1-4: >=12, Cycle 5: all respondents	C	C	C	C	C
	Other allergies	CC	Cycle 1: >=12. Cycle 2-5: All respondents	C	C	C	C	C
	Sinusitis	CC	>=12	C	C	C		
	Stomach or intestinal ulcers	CC	>=12	C	C	C	C	C
	Thyroid condition	CC	>=12		C	C	C	C
	Urinary incontinence	CC	>=12	C	C	C	C	C

C: Core, **F**: Focus, **HPS**: Health Promotion Survey, **AS**: Asthma Supplement, **FIS**: Food Insecurity Supplement, **PBI**: Provincial Buy-in, **K**: Kids' questions, **OBI**: Other Buy-ins.

Theme	Subject	Variable	Universe	1994-1995 (Cycle 1)	1996-1997 (Cycle 2)	1998-1999 (Cycle 3)	2000-2001 (Cycle 4)	2002-2003 (Cycle 5)
Drug/medication use		DG	>=12	C	C	C	C	C
Family medical history		FH	>=18			F		
Food insecurity		IN FI	Cycle 2: all respondents. Cycle 3: All respondents who were identified as food unsecured in the Household NPHS supplemental screening questions		HPS	FIS		
General health status	Self-evaluation	GH	All respondents	C	C	C	C	C
Health care	Emergency services	ES	>=12		F			
	Utilization	SV HC	Core: All respondents. HPS: >=20. PBI: >=12 Alberta	C, HPS	C, PBI	C	C	C
Health information	Health information	A HI	HPS: >=12, PBI: Alberta Ages 12 to 17	HPS	PBI			
Health status	Dexterity	HS	>=4	C	C	C	C	C
	Feelings	HS	>=4	C	C	C	C	C
	Hearing	HS	>=4	C	C	C	C	C
	Memory	HS	>=4	C	C	C	C	C
	Mobility	HS	>=4	C	C	C	C	C
	Pain and Discomfort	HS	>=4	C	C	C	C	C
	Speech	HS	>=4	C	C	C	C	C
	Thinking	HS	>=4	C	C	C	C	C
	Vision	HS	>=4	C	C	C	C	C
Height		HW	All respondents	C	C	C	C	C
Home care		HC	>=18	C	C	C	C	C

C: Core, **F**: Focus, **HPS**: Health Promotion Survey, **AS**: Asthma Supplement, **FIS**: Food Insecurity Supplement, **PBI**: Provincial Buy-in, **K**: Kids' questions, **OBI**: Other Buy-ins.

Theme	Subject	Variable	Universe	1994-1995 (Cycle 1)	1996-1997 (Cycle 2)	1998-1999 (Cycle 3)	2000-2001 (Cycle 4)	2002-2003 (Cycle 5)
Injuries	Injuries	IJ	All respondents	C	C	C	C	C
	Injury prevention	INJ	>=12	HPS				
Insurance		DV EX DG IS	Cycle 2: >=12. Cycle 3-5: All respondents		F	C	C	C
Mental health		MH	>=12	C	C	C	C	C
Nutrition	Eating habits	HW	>=12	HPS	HPS			
	Food choice	NU	Cycle 1- 2: >=12. Cycle 3 and 5: >=15	HPS	HPS	F		F
	Fruit & vegetable consumption	FV	>=15					F
	Self-perceived eating habits	GH	All respondents					C
	Supplement use	NU	>=15			F		F
Physical activities		PA	>=12	C	C	C	C	C
Pregnancy		HW PH	Females Ages 15 to 49	C	C, HPS	C	C	C
Preventive health	Blood pressure check	BP, PH	>=12	C	C	C	C	C
	Breast examination	WH	Females >=18		C			
	Breast self-examination	WH	Females >=18		HPS			
	Dental visits	DV	>=12		F			
	Eye examinations	EX	>=12		F			
	Flu shots	FS	>=12		F			
	Health improvement practices	GH A	>=12	HPS	HPS			
	HIV	SH HV	Cycle 1: Ages 15 to 45. Cycle 2: >=18	HPS	HPS			
	Mammography	WH PH	Females >=35	C	C	C	C	C
	PAP smear test	WH PH	Females >=18	C	C	C	C	C
	Physical check-ups	PC	>=12		F			

C: Core, **F**: Focus, **HPS**: Health Promotion Survey, **AS**: Asthma Supplement, **FIS**: Food Insecurity Supplement, **PBI**: Provincial Buy-in, **K**: Kids' questions, **OBI**: Other Buy-ins.

Theme	Subject	Variable	Universe	1994-1995 (Cycle 1)	1996-1997 (Cycle 2)	1998-1999 (Cycle 3)	2000-2001 (Cycle 4)	2002-2003 (Cycle 5)
Repetitive strain		RP	>=12		C	C	C	C
Residence history		RH MB	>=12					OBI
Restriction of activities		RA	Cycle 1-2: >=12. Cycle 3-5: All respondents	C	C	C	C	C
Road safety - Driving under influence		RS	>=12		HPS			
Self-care		SC	>=18			F		
Sexual health		SH	Cycle 1: Ages 15-45. Cycle 2: Ages 15-59. PBI: Alberta	HPS	HPS, PBI			
Sleep		SL	>=12					F
Smoking	Attitudes	SM	>=12		HPS			
	Awareness, attitudes, quitting	SM	>=12	HPS				
	During pregnancy	SM	Cycle 2: Female ages 15 to 49		HPS			
	Smoking	SM GH	>=12	C	C, HPS	C	C	C
	Tobacco alternatives	TA	>=12			HPS		
Social support		SS	>=12. Cycle 2-PBI: Alberta	C	C , HPS, PBI	C	C	C
Socio-demographic characteristics	Country of birth	SD	All respondents	C	C	C		
	Education	ED	>=12	C	C	C	C	C
	Ethnicity	SD	All respondents	C	C	C	C	
	Income	IN	All respondents	C	C	C	C	C
	Labour force	LF/LS	Cycle 1: >=15. Cycle 2-5: ages 15 to 75	C	C	C	C	C
	Language	SD	All respondents	C	C	C	C	C
	Race	SD	All respondents	C	C	C	C	
	Year of immigration	SD	All respondents	C	C	C		

C: Core, **F**: Focus, **HPS**: Health Promotion Survey, **AS**: Asthma Supplement, **FIS**: Food Insecurity Supplement, **PBI**: Provincial Buy-in, **K**: Kids' questions, **OBI**: Other Buy-ins.

Theme	Subject	Variable	Universe	1994-1995 (Cycle 1)	1996-1997 (Cycle 2)	1998-1999 (Cycle 3)	2000-2001 (Cycle 4)	2002-2003 (Cycle 5)
Stress	Childhood and adult stressors (traumas)	ST	Cycle 1: >=18. Cycle 4: Respondents aged less than 18 in Cycle 1 and aged 18 and over in Cycle 4	F			F	
	Coping	CO	Cycle 1: >=18 Non-Proxy Manitoba and Alberta only. Cycle 2: Alberta >=18	PBI	PBI			
	Mastery	PY	Cycle 1-5: >=12	F			F	F
	Ongoing problems	ST	>=18	F			F	F
	Recent life events	ST	>=18	F			F	
	Self-esteem	PY	>=12	F			F	
	Self-perceived	ST	Cycle 4: >=18. Cycle 5: >=12				C	C
	Sense of coherence	PY	>=18	F		F		
	Work stress	ST	Cycle 1: >=15. Cycle 4-5: ages 15 to 75	F			F	F
Two week disability		TW	All respondents	C	C	C		
UV exposure and tanning		TU	>=12. Cycle 2 - PBI: Alberta		PBI		C	C
Violence and personal safety (AB)		VS	>=12 Alberta		PBI			
Weight		HW	All respondents	C	C	C	C	C
Women's health		WH PH	Females >=18	C	C	C	C	C

C: Core, **F**: Focus, **HPS**: Health Promotion Survey, **AS**: Asthma Supplement, **FIS**: Food Insecurity Supplement, **PBI**: Provincial Buy-in, **K**: Kids' questions, **OBI**: Other Buy-ins.

Appendix B:	NPHS Household Component, Changes to the questionnaire for Cycle 5 (2002-2003)

1. Introduction

Appendix B describes the changes between the Cycle 4 (2000-2001) and Cycle 5 (2002-2003) questionnaires. Follow-up questions have been modified, and some "static" questions have been dropped (Ethnicity and Race) in Cycle 5. Improvements have been made to core content and wording. As always for each cycle, there are focus questions. In Cycle 5, stress questions from Cycle 1 and Cycle 4 are being repeated while others have been dropped. Sleep and Body Image focus modules are also new in Cycle 5. Nutrition, Alcohol Dependency and Alcohol supplement focus questions that were in Cycle 2 and 3 were also integrated. Some questions on the diagnosis and management of certain chronic conditions have been re-worded or dropped. Cycle 5 Buy-in questions were on Residence History. For more details, consult the respective questionnaires.

2. Changes to questionnaire structure

As only the longitudinal respondents are surveyed, the order of the questionnaire remained the same as in Cycle 4. Focus and buy-in contents for Cycle 5 are incorporated into the most appropriate sections of core content, or inserted as separate sections.

3. Changes to core content

In the following description, external question names from the questionnaires are used. Some internal question names may have been renamed to ensure consistency throughout the questionnaire due to the deletion or addition of questions. External variable names will be used for the dissemination of the master, share, and public use files. External variable names are created using the variable naming convention.

Household Record Variables

- Addition / revision: question on type of dwelling (DHC2_DWE) - new wording in response category 2 "Semi-detached or double (side-by-side)" changed to "Double (semi-detached)".
- Addition / revision: question on information source (AM32_SRC) - clarification of wording.

General Health (GH)

- Revision of universe statement (GH_C2) - changed from "If age < 18" to "If age < 12".
- Addition: new question on self-perceived eating habits (GHC2_4).

Preventive Health (PH)

- Addition / revision: question on hysterectomy (WHC2_5) - clarification in wording of question.

Health Care Utilization (HC)

- Addition: change to the wording in the universe statement (HC_C02).
- Addition / revision: new wording of question on consulted alternative health care provider (HCC2_4).
- Addition / revision: new wording of question on received home care services (HCC2_9).
- Addition / revision: new wording of question on nursing care service received (HCC2_10A).
- Addition: new question on home care service received not covered by government (HCC2_11).
- Addition: new question on who provided home care service (HCC2_12A to HCC2_12F) - Mark all that apply question.
- Addition: new question on type of home care services by whom provided (HCC2_3AA to HCC2_3FH) - Mark all that apply question.

Restriction of activities (RA)

- Addition / revision: new wording of question on new activity restriction or worsening of old one (RAC2_2A).
- Addition / revision: new wording of question on disappearance or improvement of activity restriction (RAC2_2B).
- Addition / revision: clarification in wording of the universe statement (RA_C5).
- Addition: new question on help needed going outdoors in any weather (RAC2_6G).

Chronic conditions (CC)

The Chronic conditions module is all Core content in Cycle 5 as opposed to Cycle 4 where part was Core and part was Focus content.

Food Allergies

- Deletion: question on age first diagnosed with food allergies (CCC0_A3).

Asthma

- Revision: question on month and year diagnosed with asthma (CCC2_C3M and CCC2_C3Y) - replaces question on age first diagnosed with asthma (CCC0_C3).
- Addition: new question on had asthma prior to last interview (CCC2_C4).

Fibromyalgia

- Revision: question on month and year diagnosed with fibromyalgia (CCC2_X3M and CCC2_X3Y) - replaces question on age first diagnosed with fibromyalgia (CCC0_X3).
- Addition: new question on had fibromyalgia prior to last interview (CCC2_X4).
- Addition: new question on last interview reported had fibromyalgia, but was not this time (CCC2_X1).
- Addition: new question on month and year fibromyalgia disappeared (CCC2_X2M and CCC2_X2Y).
- Addition: new question on received any treatment or medication for fibromyalgia (CCC2_X5).
- Addition: new question on kind of treatment or medication received for fibromyalgia (CCC2_X6A to CCC2_X6D) - Mark all that apply question.

Arthritis or Rheumatism excluding Fibromyalgia

- Revision: question on month and year diagnosed with arthritis or rheumatism excluding fibromyalgia (CCC2_D3M and CCC2_D3Y) - replaces question on age first diagnosed with arthritis or rheumatism (CCC0_D3).
- Addition: new question on had arthritis or rheumatism excluding fibromyalgia prior to last interview (CCC2_D4).
- Addition: new question on last interview reported had arthritis or rheumatism excluding fibromyalgia, but was not this time (CCC2_D1).
- Addition: new question on month and year arthritis or rheumatism disappeared (CCC2_D2M and CCC2_D2Y).
- Addition: question on kind of arthritis or rheumatism (CCC2_D11) - now part of core content, no longer a focus question (CC_0_D1).
- Addition / revision: question on received treatment for arthritis or rheumatism (CCC2_D5) - now part of core content, no longer a focus question (CC_0_D3) - new wording and removal of part of question.
- Deletion: question on had surgery for arthritis or rheumatism (CC_0_D2).
- Deletion: questions on type of surgery for arthritis or rheumatism (CC_0_D2A to CC_0_D2D) - Mark all that apply question.
- Deletion: questions on type of treatment for arthritis or rheumatism (CC_0_D3A to CC_0_D3H) - Mark all that apply question.
- Deletion: questions on type of drug for arthritis or rheumatism (CC_0_D4A to CC_0_D4C) - Mark all that apply question.
- Addition: new question on kind of treatment or medication for arthritis or rheumatism (CCC2_D6A to CCC2_D6D) - Mark all that apply question.

Back Problems

- Deletion: question on age first diagnosed with back problems (CCC0_E3).

High Blood Pressure

- Revision: question on month and year diagnosed with high blood pressure (CCC2_F3M and CCC2_F3Y) - replaces question on age first diagnosed with high blood pressure (CCC0_F3).
- Addition: new question on had high blood pressure prior to last interview (CCC2_F4).
- Revision: change of condition and wording in statement (CC_C073).

Migraine Headaches

- Deletion: revision of universe statement (CC_C081) - changed from "If age < 12" to no condition.
- Revision: question on month and year diagnosed with migraine headaches (CCC2_G3M and CCC2_G3Y) - replaces question on age first diagnosed with migraine headaches (CCC0_G3).
- Addition: new question on had migraine headaches prior to last interview (CCC2_G4).

Chronic Bronchitis or Emphysema

- Addition: Universe statement (CC_C091) containing "If age < 12".
- Deletion: question on age first diagnosed with chronic bronchitis or emphysema (CCC0_H3).

Diabetes

- Deletion: revision of universe statement (CC_C101) - changed from "If age < 12" to no condition.
- Revision: question on month and year diagnosed with diabetes (CCC2_J3M and CCC2_J3Y) - replaces question on age first diagnosed with diabetes (CCC0_J3).
- Addition: new question on had diabetes prior to last interview (CCC2_J4).
- Deletion: question on diabetes diagnosed when pregnant (CC_0_J3A).
- Deletion: question on diabetes diagnosed other than when pregnant (CC_0_J3B).
- Deletion: question on length of time started with insulin after diagnosed with diabetes (CC_0_J3C).
- Deletion: question on takes insulin daily (CC_0_J4).
- Deletion: question on number of times per day takes insulin (CC_0_J4A).
- Deletion: question on number of units per day takes insulin (CC_0_J4B).
- Deletion: question on has taken a course on managing diabetes (CC_0_J5A).
- Deletion: question on someone in household has taken a course on managing diabetes (CC_0_J5B).
- Deletion: question on has been given information on diet from a health professional on managing diabetes (CC_0_J5C).
- Deletion: question on types of health professional on managing diabetes (CC_0_J6A to CC_0_J6E) - Mark all that apply question.

- Deletion: question on has been taught how to verify sugar (CC_0_J7A).
- Deletion: question on household member has been taught how to verify sugar (CC_0_J7B).
- Deletion: question on frequency of blood sugar checks (CC_0_J8A).
- Deletion: question on frequency of checks on feet (CC_0_J8B).

Epilepsy

- Revision: question on month and year diagnosed with epilepsy (CCC2_K3M and CCC2_K3Y) - replaces question on age first diagnosed with epilepsy (CCC0_K3).
- Addition: new question on had epilepsy prior to last interview (CCC2_K4).

Heart Disease

- Deletion: question on age first diagnosed with heart disease (CCC0_L3).
- Deletion: question on number of heart attacks (CC_0_L1B).
- Deletion: question on age at first heart attack (CC_0_L1C).
- Deletion: question on age at most recent heart attack (CC_0_L1D).
- Deletion: question on ever been admitted due to heart attack (CC_0_L1E).
- Deletion: questions on treatment received for heart attack (CC_0_L4A to CC_0_L4E) - Mark all that apply question.
- Deletion: question on ever been referred to a cardiac rehabilitation program (CC_0_L5A).
- Deletion: question on has attended a cardiac rehabilitation program (CC_0_L5B).
- Deletion: question on completed a cardiac rehabilitation program (CC_0_L5C).
- Addition: question on ever had a heart attack (CCC2_L1A) - now part of core content, no longer a focus question (CC_0_L1A).
- Addition: question on has angina (CCC2_L6) - now part of core content, no longer a focus question (CC_0_L6).
- Addition: question on has congestive heart failure (CCC2_L7) - now part of core content, no longer a focus question (CC_0_L7).

Cancer

- Deletion: revision of universe statement (CC_C131) - changed from "If age < 12" to no condition.
- Deletion: question on age first diagnosed with cancer (CCC0_M3).

Intestinal or Stomach Ulcers

- Revision: terms intestinal and stomach switched (CCC2_N1).
- Revision: question on month and year diagnosed with stomach or intestinal ulcers (CCC2_N3M and CCC2_N3Y) - replaces question on age first diagnosed with stomach or intestinal ulcers (CCC0_N3).
- Addition: question on had stomach or intestinal ulcers prior to last interview (CCC2_N4).

Effects of a stroke

- Revision: question on month and year diagnosed with suffering from the effects of a stroke (CCC2_O3M and CCC2_O3Y) - replaces question on age first diagnosed with suffering from the effects of a stroke (CCC0_O3.
- Addition: new question on suffered from the effects of a stroke prior to last interview (CCC2_O4).

Urinary Incontinence

- Deletion: question on age first diagnosed suffering from urinary incontinence (CCC0_P3).

Bowel Disorder

- Deletion: question on age first diagnosed with a bowel disorder (CCC0_Q3).

Alzheimer's Disease or other Dementia

- Addition / revision: question on has Alzheimer's disease or other dementia (CCC2_1R) - clarification of word "senility".
- Deletion: question on age first diagnosed with Alzheimer's Disease or other dementia (CCC0_R3).

Cataracts

- Deletion: question on age first diagnosed with cataracts (CCC0_S3).

Glaucoma

- Deletion: question on age first diagnosed with glaucoma (CCC0_T3).

Thyroid Condition

- Deletion: question on age first diagnosed with a thyroid condition (CCC0_U3).

Insurance (IS)

- No change.

Health Status (HS)

- No change.

Physical Activities (PA)

- Addition / revision: question on activity in the past 3 months (PAC2_1I), new wording for response category 13 "Downhill skiing or snowboarding".

UV Exposure (TU)

- No change.

Repetitive Strain (RP)

- No change.

Injuries (IJ)

- Addition / revision: question on received medical attention for injury (IJC2_11) - change to wording in sentence.

Medication Use (DG)

- Addition / revision: question on medication use (DGC2_1B) - new wording for examples of tranquilizers "such as Valium or Ativan".
- Addition / revision: question on medication use (DGC2_1C) - new wording for examples of diet pills "such as Ponderal, Dexatrim or Fastin" .
- Addition / revision: question on medication use (DGC2_1F) - new wording for examples of allergy medicine "such as Reactine or Allegra" .
- Addition / revision: question on medication use (DGC2_1P) - new wording for examples of sleeping pills "such as Imovane, Nytol or Starnoc" .
- Addition / revision: question on use of another health product (DGC2_5AA to DGC2_5LA) - question wording changed to "Did you use another health product?" - up to 12 health products.
- Addition / revision: flag for long answer collected for exact name of next product used (DGC2F5B to DGC2F5L) changed to "What is the exact name of this product?" - up to 11 health products.

Smoking (SM)

- Addition / revision: question on previously smoked, but this time does not, ever smoked (SMC2_4B).
- Addition: new question on cigarette brand usually smoked (SMC2C8B).

Alcohol (AL)

- No change.

Mental Health(MH)

- Addition / revision: question on has consulted a health professional about mental health (MHC2_1K) - slight change of wording in question.

Social Support (SS)

- No change.

Socio-demographic Characteristics (SD)

Ethnicity

- Deletion: question on ethnicity (static question previously asked) (SDC0_4A to SDC0_4S) - Mark all that apply question.

Race

- Deletion: question on race (static question previously asked) (SDC0_7A to SDC0_7L) - Mark all that apply question.

Language

- No change.

Education(ED)

- No change.

Labour Force (LF / LS)

- No change.

Income (IN)

- No change.

Provincial Health Number and Administration (AM)

- Addition / revision: question on permission to link (AM62_LNK) - new wording in question.
- Addition / revision: question on agreement to share information (AM62_SHA) - new wording in question.
- Addition / revision: question on correction to name (AM62_15) - new wording in question.

4. Changes to focus content

New focus sections on Sleep and Residence History were introduced. Cycle 3 Nutrition module questions were included. As well, certain questions asked in Cycle 1 and Cycle 4 Stress module were repeated (Ongoing Problems, Work Stress and Mastery). Cycle 2 Alcohol Opinion supplement and Alcohol Dependence modules were also repeated for Cycle 5.

For Cycle 5 the following focus content has been included:

Sleep (SL)

- Addition: 4 new selected questions taken from CCHS (SL_2_1 to SL_2_4).

Height and Weight (HW)

Body Image

- Addition: 5 new questions (HW_2_4 to HW_2_8) and one condition (BI_C1).

Nutrition (NU)

- Questions on food choice and supplement use are a repeat of the questions asked in Cycle 3 with the following changes:

Food choice

- Deletion: question on chooses or avoids foods to maintain or improve health (NU_8_1B).
- Deletion: question on chooses or avoids foods because concerned about high blood pressure (NU_8_1F).
- Deletion: question on chooses or avoids foods because concerned about diabetes (NU_8_1G).
- Deletion: question on chooses foods for iron content (NU_8_2D).
- Deletion: question on chooses foods for other vitamins or minerals (NU_8_2E).
- Deletion: question on avoids foods for sugar content (NU_8_3E).
- Deletion: question on avoids foods for iron content (NU_8_3F).

Supplement use

- No change.

Fruit and Vegetable consumption - FV

- Addition - 12 new questions taken from CCHS (FV_2_Q1 to FV_2_Q6) - Each question contains number of times and reporting period.

Stress (ST)

These questions are a repeat of the questions asked in Cycle 1 and Cycle 4 with the following changes:

Ongoing problems

- Addition: new wording in question (ST_QINT1B).

Recent life events

- Deletion: questions on recent stress (ST_0_R1 to ST_0_R10).

Childhood and adult stressors ("traumas")

- Deletion: questions on traumas (ST_0_T1 to ST_0_T7).

Work stress

- No change.

Self-esteem

- Deletion: questions on self-esteem (PY_0_E1A to PY_0_E1F).

Mastery

- No change.

Alcohol (AL)

Alcohol attitude supplement

- These questions (AL_2_1 to AL_2_7) are a repeat of the questions asked in Cycle 2 Alcohol supplement module for Cycle 5.

Alcohol Dependence (AD)

- These questions (AD_2_1 to AD_2_9) are a repeat of the focus questions asked in Cycle 2.

5.   Changes to buy-in content

Residence History (RH / MB)

- Addition: new module for Cycle 5; it contains 2 alternative sets of questions: one starting at current year (or year of last move) and going back to 1980; one starting in 1980 and ending in current year (or year of last move).

Appendix C: NPHS Household Component, Examples of Cycle 4 (2000-2001) Data Feedback and Follow-up Questions

Blood Pressure; Mammography; Pap Smear Test	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 and 5, 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.
Restriction of Activities	Information on whether or not the respondent had a disability in Cycle 3 was used in Cycle 5. If the status changed, an explanation of that change was probed.
Chronic Conditions	For all respondents, selected chronic conditions (asthma, fibromyalgia, arthritis, high blood pressure, migraine headaches, diabetes, epilepsy, stomach or intestinal ulcers and the effects of a stroke) were fed back in an attempt to help explain change. If it was a newly acquired condition, the date of onset for the condition was acquired.
Smoking	If a daily smoker had reported the age at which he or she started smoking daily during last interview, that response was fed back in Cycle 5. For the occasional smoker or non-smoker in Cycle 5 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.
Socio-demographic Characteristics	For all respondents, a flag indicating that country of birth had been collected was input again. Language first learned and still spoken was asked again because it can change over time.

Education	For all respondents, a flag indicating the highest level of education was re-input. Screening questions determined if the respondent was currently attending a learning institution between cycles. If so, educational attainment was collected anew.
Labour Force	For all respondents, the employer name, type of industry and duties of the main job in Cycle 4 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.