STATISTICS CANADA

NATIONAL POPULATION HEALTH SURVEY

HEALTH INSTITUTIONS COMPONENT

CYCLE 5 (2002-2003)

LONGITUDINAL DOCUMENTATION

January 2007

Statistics Canada Statistique Canada

Canada

NOTE TO USERS

Starting with the collection of Cycle 3 data in 1998-1999, the National Population Health Survey (NPHS) – Health Institutions Component was strictly longitudinal in nature.

To provide greater flexibility to users, one microdata master file is being issued for NPHS Health Institutions component Cycle 5. This file includes all 2,287 NPHS Health Institutions panel members, notwithstanding their response patterns from previous cycles. The master file has three subsets of respondents with corresponding sampling weights and a flag to make their identification easier.

The NPHS Cycle 5 longitudinal documentation provides a wide range of information on the survey: objectives, survey content, sample design, collection, processing, data quality, weighting procedures, tabulation's guidelines and data access. Chapter 12 gives more details on the various subsets of respondents and their associated weights.

This guide is also intended for users of the share file, i.e. provincial ministries of health, Health Canada and the Public Health Agency of 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.

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

Cycle 1 = 1994-1995
Cycle 2 = 1996-1997
Cycle 3 = 1998-1999
Cycle 4 = 2000-2001
Cycle 5 = 2002-2003

Table of Contents

List of tables

1. Introduction

The National Population Health Survey (NPHS) is designed to collect information on the health of the Canadian population and related socio-demographic information. The Health Institutions component of the NPHS is the first national longitudinal survey of residents of Canadian health care facilities. The first cycle of data collection took place in 1994-1995. The NPHS fulfilled both cross-sectional and longitudinal needs during its first two cycles, and then with Cycle 3 (1998-1999) NPHS Health Institutions became strictly a longitudinal survey.

The NPHS Health Institutions component includes residents of long-term (more than 6 months) health care institutions in all provinces, but excluded the territories, Indian Reserves and Canadian Forces Bases. The institutions considered were long term health care institutions with at least 4 beds and where residents cannot be autonomous. The Health Institutions component of the NPHS has completed five release 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).

After 5 cycles of collection (1994/1995 to 2002/2003), the institutions component ends due to the large number of deaths in the sample. Respondents from the household component who move to an institution continue to be followed and remain part of the household sample.

The Cycle 5 NPHS Health Institutions component collected in-depth information on the health of the longitudinal respondent who was randomly selected in Cycle 1. As in Cycle 4, the collection of data from persons in the institution sample who moved to households was done by the Health Institutions component, rather than the Household component. Questions for household residents were separated from those of the institutions residents with over 90% of the questions identical.

This document has been produced to facilitate the use of the Cycle 5 (2002-2003) Longitudinal Master and Share Files from the NPHS Health Institutions 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:

Data Access and Information Services
Health Statistics Division

Information request, Access (Research Data Centres, Remote Access)
Tel: 1-613-951-1653, E-mail: nphs-ensp@statcan.ca, Fax: 1-613-951-0792

Custom tabulations/general data support
Tel: 1-613-951-1746, E-mail: hd-ds@statcan.ca, Fax: 1-613-951-0792

2. Background

In the fall of 1991, the National Health Information Council (NHIC) recommended that an on-going national survey of population health be conducted. This recommendation was based on consideration of the economic and fiscal pressures on the health care system and the commensurate requirement for information 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 of factors having an impact on health.

Beginning in April 1992, Statistics Canada received funding for the 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.

A special component covering residents of health institutions was undertaken because this population was rarely covered by national surveys and likely had health characteristics different from those of the general population.

3. Objectives

The objectives of the NPHS are to:

- aid in public policy development 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, and environmental correlates of health;
- increase the understanding of the relationship between health status and health care utilization;
- provide information on a panel of people followed over time, to reflect the dynamic process of health and illness and determine the factors affecting institutionalization;
- provide the provinces and territories and other clients with a health survey capacity that will allow supplementation of content or sample;
- allow the possibility of linking survey data to administrative data that are collected routinely, such as vital statistics, environmental measures, community variables, and health services utilization.

4. Survey Content

4.1. Criteria

The content of the NPHS Health Institutions component was selected according to the following criteria:

1) The survey should collect information on the health status of the Canadian population residing in health institutions.
2) The data collected should be comparable to that of the household population whenever possible.
3) The survey should increase the understanding of conditions related to institutionalization.
4) Information provided should permit the study, over time, of the transitions from households to institutions and vice versa.
5) The survey should produce national level data.

Respondents were randomly chosen from selected health care institutions. Two questionnaires were used to collect data; the Institution Control Form (ICF) and the Respondent Questionnaire (RQ). The ICF asked about the policies of the Institution and the RQ included questions on health status, risk factors, social support, contact with health care providers, and demographic and social-economic status. For example, health status was measured through questions on self-perception of health, functional ability, chronic conditions and activity restriction. Behavioural risk factors included smoking and alcohol use. The level of social support was assessed by the frequency of contact with friends and relatives inside and outside the institution. Demographic and socio-economic information included age, sex, education, ethnicity and personal income.

4.2. Content Revisions for Cycle 5 (2002-2003)

There were only a few content changes for Cycle 5. The following items were modified:

- The "prefill" for year for Date of Interview on the front page of the respondent questionnaire was changed to 2003.
- The permission to link question was re-worded.
- The permission to share question was re-worded.
- The final status codes on the front cover of the respondent questionnaire were changed to reflect the standard codes being used by all Statistics Canada Surveys.

5. Sample Design

The target population of the 1994-1995 NPHS Health Institutions component consisted of the residents of long-term (more than 6 months) health care institutions in all provinces, but excluded the territories, Indian Reserves and Canadian Forces Bases. Health care institutions considered were the long term institutions with at least 4 beds and where residents cannot be autonomous. Institutions that are not part of the health care system, such as correctional facilities, prisons, young offender facilities, orphanages and religious institutions are not included in the survey frame of health care institutions.

The longitudinal sample of the NPHS Health Institutions component consists of all longitudinal respondents chosen in Cycle 1 (1994-1995). The Cycle 1 sample size was set at 2,600 residents. Assuming a response rate of 85%, this sample size would be sufficient to calculate national estimates with a coefficient of variation (CV) of 10% for characteristics occurring in a minimum of 10% of the population.

After Cycle 1 collection, this sample consisted of 2,287 persons living in a health institution in 1994-1995.

The selection of the respondents was done in two stages. First, health institutions were selected, then residents were selected within these institutions.

5.1. Cycle 1 (1994-1995) Stratification and Selection of Health Institutions

A list of in-scope health institutions was created (long-term, at least 4 beds and residents not autonomous). This list was initially stratified by geographic region (geographic strata) and subsequently by the type of institution (characteristic strata) and number of beds (size strata).

There were five geographic strata; the Atlantic provinces, Quebec, Ontario, the Prairie provinces and British Columbia. Within each geographic stratum three characteristic strata were defined:

Institutions for the Aged	including residential care facilities for the aged and extended/chronic care hospitals.
Cognitive Institutions	including residential care facilities for emotionally disturbed children, psychiatrically disabled and developmentally delayed people and psychiatric hospitals.
Other Rehabilitative Institutions	including rehabilitation, pediatric and other speciality hospitals, general hospitals with long-term units as well as residential care facilities for people with physical disabilities.

Within each of these geographic/characteristic strata, the institutions were grouped into size strata by grouping facilities with a similar number of beds. The number of size strata created depended on the total number of beds in the geographic/characteristic strata. Once the number of size strata was determined, the boundaries for the different size strata were fixed using the $Cum \sqrt{f(y)}$ rule where *f(y)* was the number of beds.

In Cycle 1, the number of institutions selected from a size stratum depended on the amount of sample allocated to the stratum (see Section 5.2) and the size of the institutions within the stratum. In strata comprised of larger institutions, a larger sample of residents was selected from each institution. This reduced the total number of institutions visited. Once the number of institutions to be selected from each size stratum was determined, a systematic sample of institutions was taken from the stratum list with the probability of selection proportional to size (PPS). Size was determined by the number of long-term beds.

It was possible that the listing indicated a head office for several smaller institutions. In this case, a listing of all of the institutions under this head office was obtained and two were selected: the largest (in terms of beds) and another randomly selected using PPS sampling.

5.2. Cycle 1 (1994-1995) Selection of Residents

Once the institution had been selected, residents of these institutions were selected. The total sample of 2,600 residents was proportionally allocated to each of the size strata based on the number of beds in each stratum. The sample was increased to thirty residents when a size stratum had an initial sample size of less than thirty residents.

After Cycle 1 collection, this sample consisted of 2,287 persons living in a health institution in 1994-1995.

5.3. Longitudinal Sample

The longitudinal sample, also called the longitudinal panel or simply the panel, is composed of the 2,287 persons that were selected in Cycle 1 and had partially or fully completed the questionnaire in Cycle 1. This panel was surveyed in Cycles 2, 3, 4 and 5. The institutions component ends with Cycle 5, due to the large number of deaths in the panel.

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 (2,287) for all cycles.

The number of people answering the survey slightly decreases from one cycle to the next due to attrition caused by non-respondents (for example refusals or individuals that were untraceable). Despite the attrition, the longitudinal sample is still representative of the

1994-1995 target population. The attrition being very small (see Section 8.2), it should not lead to large increases in the variance of estimates. It should be noted that panel members who died or who moved to a household are still considered part of the institutional component of the longitudinal sample. The panel members who died 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 in 1994-1995 and shows the size of the longitudinal full subset (see section 7.5), for each cycle. The table also contains the number of deceased persons within this subset.

Table 5.A: Size of the longitudinal full subset, for each cycle

Cycle	Longitudinal Full Subset	***Number of deceased*** *(Within the longitudinal full subset)*
1	2,287	--
2	2,192	721
3	2,178	1,250
4	2,143	1,524
5	2,131	1,708

6. Data Collection

6.1. Questionnaire Design and Data Collection

The NPHS Health Institutions component questions were designed to be conducted by personal interview using paper and pencil. Telephone interviews were acceptable when a proxy respondent could not be contacted in person. The administrator of the institution or a contact within the institution determined which of the selected residents required a proxy interview. This decision was based on the selected respondents' health status. The proxy respondent could be a relative, a staff member or a volunteer at the institution. Proxy respondents completed 72% of the interviews (of the proxy interviews, 34% were done by relatives of the resident). A staff member from the institution provided information on each selected resident's use of medications and their contact with health professionals.

Collection took place from April until June 2003. Statistics Canada interviewers conducted the interviews. At the beginning, all institutions were contacted by telephone by an interviewer to arrange a meeting between the interviewer and the administrator or contact person from the institution. During this liaison visit, the interviewer administered a short questionnaire on the policies of the institution. The residents requiring proxy interviews were also determined at this time. The name and telephone number of the next-of-kin were obtained in these cases. The next-of-kin was then phoned and given the option to complete the interview primarily themselves or have it completed by a knowledgeable institutional staff member.

All interviewers were under the supervision of senior interviewers. The senior interviewers were responsible for ensuring that interviewers were familiar with the concepts and procedures of the survey. They periodically monitored interviewers and reviewed their completed documents. The senior interviewers were, in turn, under the supervision of project managers, located in each of the Statistics Canada Regional Offices.

6.2. Non-Response to the NPHS

Interviewers were instructed to make all reasonable attempts to obtain interviews with selected residents or proxy respondents. Refusals at the institutional level were followed-up by senior interviewers, project managers or by other interviewers to try to convince the institution to participate in the survey, with the result of having no refusals at the institutional level in Cycle 5.

7. Data Processing

7.1. Data Capture and Editing

After completing an interview, the interviewer reviewed the questionnaire to verify that the correct flow of questions was followed during the interview. Further editing was done at the Regional Offices to check for completeness, legibility and consistency of entries on the questionnaire. This allowed for immediate follow-up with the respondent.

The respondent questionnaire and Institution Control Form were captured at the Head Office using EP90 (Entry Point 90). The programs written for the data capture prevented most out-of-range values from being entered. All captured information, excluding comments, was 100% verified.

After data capture, questionnaire data flows were verified and consistency edits between certain fields were performed. With the exception of the Health Utility Index (HUI3), no imputation was performed (see Section 8.4).

7.2. Coding

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). For Cycle 5, conditions or health problems causing activity restrictions were also coded based on the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10). Both codes (ICD-9 and ICD-10) appear on the longitudinal file. The Cause of Death (if applicable) has also been coded to both ICD-9 and ICD-10.

The drug classification is based on the Anatomical Therapeutic Chemical (ATC) Classification System developed by the World Health Organisation as available on the Health Canada Drug Product Database (DPD) in September 2003. A complete revision of the drug codes was done for all NPHS longitudinal respondents for Cycle 5 (2002-2003) and for all previous cycles.

7.3. Creation of Derived Variables

To facilitate data analysis, a number of variables on the file have been derived using responses to the NPHS questionnaire for respondents in health institutions. A “D” appearing in the fifth position of the variable name indicates the variable is derived. Details of how the derived variables were created can be found in the documentation on derived variables.

7.4. Estimation and Weighting

The principal 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 in Canada 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 10.

7.5. Subsets of respondents

In order to provide greater flexibility to users, a single microdata master file has been created for NPHS Cycle 5, Health Institutions. This file includes all 2,287 NPHS panel members, notwithstanding their response patterns from previous cycles. Within the master file, three subsets of respondents have been created along with corresponding sampling weights and a flag to make their identification easier. Refer to Chapter 10 for more information regarding the calculation of each subset's sampling weight and to Section 12.1 for the use of longitudinal weights. Table 7.A provides a description of the three subsets of respondents based on the type of response.

Table 7.A: Subsets of respondents

Subset of respondents	Type de response	Flags	Number of respondents
Longitudinal Square	Complete panel: all panel members regardless of their response pattern in Cycles 1, 2, 3, 4 and 5.	None, all records	2,287
Longitudinal Full	All panel members with a complete response (Full) in Cycles 1, 2, 3, 4 and 5.	**WFI2LF**	2,131
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.	**WFI2SLF**	2,067

7.6. Definition of the longitudinal response pattern

In each cycle, each member of the panel is assigned a response status. Depending on the outcome of the interview, a person is assigned one of the following four statuses: household (1), dead (2), institutionalized (3) or non-response (5). A series of rules exist for determining each response status. A panel member who provided a complete or a partial response to the interview has a status of "institutionalized "or "household" depending if the panel member lives in an institution or a household. If the longitudinal member is not deceased, institutionalized or in a household, that person is considered a "non-response". For the NPHS Health Institutions longitudinal panel, a complete response includes status "household", "dead" and "institutionalized".

Over the cycles, the response statuses of a given respondent are concatenated into a single variable called the "longitudinal response pattern" (LONGPAT). This variable is available on the longitudinal file and can be used to obtain rapidly the response profile of a member of the panel. This variable is also used to identify different analytical subsets as described in Section 7.5.

8. Data Quality

8.1. Longitudinal Response Rate

Two separate response rates can be calculated from the longitudinal file of the NPHS Health Institutions component, the response rate for institutions and the response rate for individuals. 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 2,444 in-scope persons selected to form the panel while response rates for subsequent cycles are based on the 2 287 individuals who form the longitudinal panel.

8.1.1. Cycle 1 Response Rate (1994-1995)

Institutions Response Rate

The institutions response rate corresponds to the percentage of in-scope institutions that agreed to have the survey conducted among their residents[1]. Residents could not be interviewed without the institution's permission. The institutions response rate was calculated as follows:

$$\frac{\textit{Number of institutions selected that agreed to participate}}{\textit{Total number of institutions where panel members resided}} \times 100$$

$$= \frac{214}{224} \times 100 = 95.5\%$$

Individual Response Rate

The individual response rate corresponds to the percentage of selected residents from the responding institutions with whom an interview was conducted. This rate is calculated as follows:

$$\frac{\textit{Number of residents who participated fully or partially in an interview}}{\textit{Total number of selected residents in the participating institutions}} \times 100$$

$$= \frac{2{,}287}{2{,}444} \times 100 = 93.6\%$$

NOTE: Multiplying the two rates together does not produce a valid result because the number of residents chosen varies by institution.

[1] Residents who were part of the 8 out-of-scope institutions were excluded from the panel and from the calculations of the Cycle 1 response rates.

8.1.2. Response Rate for Cycle 2 (1996-1997), Cycle 3 (1998-1999), Cycle 4 (2000-2001) and Cycle 5 (2002-2003)

Institutions Response Rate

The institutions response rate is based on the total number of institution where panel members resided (e.g. all of the institutions that participated in the first cycle of the survey and that were still in operation and all institutions newly covered) and is calculated as follows:

$$\frac{Number\ of\ institutions\ selected\ that\ agreed\ to\ participate}{Total\ number\ of\ institutions\ where\ panel\ members\ resided} \times 100$$

The institutions response rates were 100 % in Cycles 2, 3 and 5, and 99.3 % in Cycle 4. Those results are presented in table 8.A.

Table 8.A: Institutions Response Rate

Cycle	Number of institutions where panel members resided		Response Rate
	Agreed to participate	Total	
	(1)	***(2)***	***(1)/(2)***
2	314	314	100.0%
3	352	352	100.0%
4	287	289	99.3%
5	270	270	100.0%

Individual Response Rate

For Cycles 2, 3, 4 and 5, the individual response rates are based on the 2,287 individuals who form the longitudinal panel. Persons who have died or who have moved to a household are counted as a response for longitudinal purposes. This rate is calculated as follows:

$$\frac{Number\ of\ residents\ who\ participated\ fully\ or\ partially\ in\ an\ interview}{Total\ number\ of\ residents\ chosen\ in\ the\ participating\ institutions} \times 100$$

The individual response rates were 95.9 % in Cycle 2, 98.4 % in Cycle 3, 96.9 % in Cycle 4 and 97.7 % in Cycle 5. Those results are presented in table 8.B.

Table 8.B: Individual Response Rate

Cycle	Number of residents who participated fully or partially in an interview				Response Rate = *(1)+(2)+(3) / 2,287*
	Institutionalized *(1)*	In a household *(2)*	Deceased *(3)*	Total *(1)+(2)+(3)*	
2	1,427	44	722	2,193	95.9%
3	907	73	1,271	2,251	98.4%
4	595	50	1,572	2,217	96.9%
5	421	41	1,773	2,235	97.7%

NOTE: Multiplying the institution response rate and the individual response rate together does not produce a valid result since the number of residents chosen varies by institution.

8.2. Attrition Rate

Attrition is a loss in sample size due to non-respondents, movements out-of-scope and untraceable individuals. For the NPHS Health Institution component, attrition is very low given the high response rate. It is important to remember that deceased respondents are part of the longitudinal full subset and are considered as "respondents".

Two different attrition rates can be calculated: one showing the attrition rate observed at the end of each cycle, the other showing the cumulative attrition rate based on the original sample. Both of these rates are calculated using the number of individuals found in the Full subset of respondents (see Section 5.3).

Relevant information for calculation of attrition rates:

Number of longitudinal panel members:	2,287
Number of individuals in the Cycle 2 (1996-1997) Longitudinal Full subset	2,192
Number of individuals in the Cycle 3 (1998-1999) Longitudinal Full subset	2,178
Number of individuals in the Cycle 4 (2000-2001) Longitudinal Full subset	2,143
Number of individuals in the Cycle 5 (2000-2001) Longitudinal Full subset	2,131

Attrition rates at the end of each cycle

Cycle 2 (1996-1997): $\frac{2,287 - 2,192}{2,287} = \frac{95}{2,287} = 4.2\%$

Cycle 3 (1998-1999): $\frac{2,192 - 2,178}{} = \frac{14}{} = 0.6\%$

2,192 2,192

Cycle 4 (2000-2001): $\frac{2,178 - 2,143}{2,178} = \frac{35}{2,178} = 1.6\%$

Cycle 5 (2002-2003): $\frac{2,143 - 2,131}{2,143} = \frac{12}{2,143} = 0.6\%$

Cumulative attrition rates

Cycle 2 (1996-1997): $\frac{2,287 - 2,192}{2,287} = \frac{95}{2,287} = 4.2\%$

Cycle 3 (1998-1999): $\frac{2,287 - 2,178}{2,287} = \frac{109}{2,287} = 4.8\%$

Cycle 4 (2000-2001): $\frac{2,287 - 2,143}{2,287} = \frac{144}{2,287} = 6.3\%$

Cycle 5 (2002-2003): $\frac{2,287 - 2,131}{2,287} = \frac{156}{2,287} = 6.8\%$

8.3. Survey Errors

8.3.1. Sampling Errors

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

Estimates produced from a sampling survey include a sampling error. Good statistical techniques require that researchers provide users with some indication of the size of that error. This part of the documentation describes the *sampling error measures* that Statistics Canada normally uses and which it recommends users to adhere to when deriving estimates from this master file.

Determination of the possible size of sampling errors is based on the standard error of estimate derived from the survey results. Given the large variety of estimates that can be produced from a survey, the standard error of an estimate is usually expressed relative to the estimate to which it pertains. The resulting measure, known as the coefficient of variation (CV) of an estimate, is obtained by

dividing the standard error of the estimate (equal to the square root of the variance of the estimate) by the estimate itself, and is expressed as a percentage of the estimate.

For example, suppose that based upon the survey results, one estimates that *10.4%* of residents in in-patient health care institutions are daily cigarette smokers and that the standard error for this estimate is *0.0094*. The coefficient of variation is calculated as follows:

$$\left(\frac{0.0094}{0.104}\right) x\ 100\ \% = 9.04\ \%$$

Chapter 11 contains more details on the calculation of the variance for this survey. Please consult section 9.4 for the interpretation of the CV and the guidelines for release.

8.3.2. Non-sampling Errors

Errors not related to sampling may occur at almost every stage of a survey. Instructions may not be clear or may be misinterpreted by the interviewers, respondents may make errors answering the questions, answers may be incorrectly recorded on the questionnaire or errors may be introduced during the processing and tabulation of the data. These are all examples of *non-sampling errors*.

Over a large number of observations, randomly occurring errors will have little effect on estimates derived from the survey. However, errors occurring systematically will contribute to biases in the survey estimates. Considerable time and effort was made to reduce non-sampling errors in the survey. Quality assurance measures were implemented at each step of the data collection and processing cycle to monitor the quality of the data. These measures included the use of highly skilled interviewers, extensive training of interviewers with respect to the survey procedures and questionnaire, observation of the interviewers in order to identify any problems and adoption of procedures to minimize data collection and capture errors.

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 occurred when a respondent refused to answer a question or could not remember the information requested. Total non-response occurred when the interviewer was unable to communicate with the person responsible for answering by proxy or the respondent chosen refused to participate in the survey. In the case of the NPHS Health Institutions component, both partial and total non-response are low. Total non-response cases are handled by correcting the weight of residents who responded to the survey in order to compensate for those who did not respond (refer to Chapter 10 for more information on weighting).

8.4. Imputation

Imputation was used to derive the missing values for one variable in the NPHS Health Institutions component. The variable HSI2DHSI denotes the respondent's Health Utility Index (HUI3)[2]. This measure of overall health status assesses vision, hearing, speech, getting around (ability to move about), dexterity (movement of hands and fingers), feelings, cognitive ability (memory and thinking) and pain. The overall HUI3 rating, which can range from -0.360 to 1.000 is calculated based on responses to a series of questions on health status. However, this overall rating cannot be calculated if one or more of the answers are missing, a situation that occurs for about 10% of respondents. It was decided to use imputation for the missing values in order to calculate the HUI3 in Cycle 5 (2002-2003) of the health care institutions component. The **hot deck method** was therefore used to impute values for the missing elements in order to be able to calculate the overall HUI3 for the individuals concerned. It should be noted that the same method was used in previous cycles.

The HUI3 was calculated based on the answers to questions on the eight elements in the health status section. A partial rating was calculated for each of the elements and then further calculations were done on these partial ratings to derive the overall HUI3 rating. Imputation was at the level of the eight partial ratings rather than the questions. After imputation, the program for calculating the derived HUI3 variable was changed slightly so that it selected as entry data the eight imputed values for vision, hearing, speech, getting around, feelings, cognitive ability, dexterity and pain.

Imputation was done in two stages:

- The first stage used a deterministic imputation. In some instances, even if the person did not answer the question providing the partial rating, there was sufficient information to deduce the partial rating with certainty. Therefore, a partial rating based on this partial information was attributed in all instances where it was considered appropriate to do so.

- The second stage corresponds to a hot deck donor imputation to attribute the missing partial ratings. The nearest neighbour method was used to identify the donors. The nearest neighbour was determined by calculating a temporary HUI3, using only the partial ratings containing only valid values.

[2] For more information on the calculation of the HSI, see the derived variable documentation.

9. Guidelines for Tabulation, Analysing and Release

This section of the documentation describes the guidelines to be adhered to by users tabulating, analysing, publishing or otherwise releasing any data derived from the survey files. With the aid of these guidelines, users should be able to reproduce the figures produced by Statistics Canada and also to develop currently unpublished figures in a manner consistent with these established guidelines.

9.1. Rounding guidelines

Below are the guidelines to be followed when rounding estimates derived from the data files:

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

b) Marginal subtotals and totals of statistical tables are to be derived from their corresponding unrounded components and then rounded themselves to the nearest 100 units using normal rounding.

c) Averages, proportions, rates and percentages are to be computed from unrounded components (i.e., numerators and/or denominators) and then rounded themselves to one decimal using normal rounding. In normal rounding to a single decimal number, 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 increased by one (1).

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

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

f) Unrounded estimates are not to be published or otherwise released under any circumstances. Unrounded estimates give the impression of being much more accurate than they are in reality.

9.2. Sample Weighting Guidelines for Tabulation

The sample design used for the NPHS Health Institutions component is not self-weighted. In other words, the sampling weight is not the same for all respondents. Even when deriving simple estimates, including standard statistical tables, the user must use the appropriate sampling weight. If this is not done, estimates calculated from this file will not be deemed to be representative of the surveyed population and will not correspond to the estimates produced that may be produced by Statistics Canada.

The user should also remember that some software programs do not take weights into consideration, which prevents users from obtaining estimates that match exactly those of Statistics Canada.

9.2.1. Definitions of Estimate Categories: Categorical Versus Quantitative

Two main types of point estimates of the characteristics of the population can be derived from the data file of the NPHS Health Institutions component.

Categorical estimates:

Categorical estimates (also called estimates of an aggregate) are estimates of the number or percentage of persons who, in the surveyed population, have certain characteristics or are part of a specific category. The number of individuals who smoke daily is an example of this type of estimate.

Example of a categorical question:

SMI2_1: At the present time, do you (does . . .) smoke cigarettes daily, occasionally or not at all?

__ Daily
__ Occasionally
__ Not at all

Quantitative estimates:

Quantitative estimates are estimates of totals or of means, medians or other measures of central tendency of quantities based on some or all of the members of the surveyed population. They also explicitly include estimates of the form $\hat{Y} / \hat{X}$ where $\hat{Y}$ is an estimate of the total quantity for the surveyed population and $\hat{X}$ is

an estimate of the number of people in the surveyed population who contribute to that total quantity.

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

Example of a quantitative question:

SMI2_3: How many cigarettes do you (does . . .) smoke each day now?

|_|_| Cigarettes

9.2.2. Tabulation of Categorical Estimates

Estimates of the number of individuals with a certain characteristic can be obtained from the data file by summing the weights of all records possessing the characteristic(s) of interest. Proportions and ratios of the form $\hat{Y} / \hat{X}$ are obtained by:

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

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

c) dividing the numerator estimate (obtained in "a") by the denominator estimate (obtained in "b").

9.2.3. Tabulation of Quantitative Estimates

Estimates of quantities can be obtained by multiplying the value of the variable of interest by the weight of each record, then adding this quantity for all of the records concerned. For example, to obtain an estimate of the *total* number of cigarettes smoked per day by individuals who smoke daily, the value reported in question SMI2_3 is multiplied by the weight for the record (WTI2LF) and then this value is summed over all records with a response of "daily" to SMI2_1.

To obtain a weighted average expressed in the form $\hat{Y} / \hat{X}$, the numerator ($\hat{Y}$) was calculated as a quantitative estimate and the denominator ($\hat{X}$) as a categorical estimate. For example, the *average* number of cigarettes smoked per day by individuals who smoke daily can be obtained by:

a) estimating the total number of cigarettes smoked per day by individuals who smoke daily using the above method;

b) estimating the number of individuals who smoke daily by summing the weights of all records in which the response to question SMI2_1 is "daily";

c) dividing the numerator estimate (obtained in "a") by the denominator estimate (calculated in "b").

9.3. Guidelines for Statistical Analysis

The NPHS Health Institutions component is based on a two-stage sample design where the institutions were selected without replacement. Using data from this type of survey presents difficulties for analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

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. While in many cases the estimates produced by the packages are correct, the variances that are calculated are almost meaningless.

To calculate the variance of an estimate obtained with the NPHS data, it is recommended to use the bootstrap method along with the Bootvar program provided with the data (see Chapter 11).

With many statistical packages, it is possible for many analysis techniques (for example linear regression, logistic regression, analysis of variance), to 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 will not allow for the stratification 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 which is equal to the original weight divided by the average of the original weights for the sampled units (people) contributing to the estimator in question.

9.4. Release Guidelines

Before releasing and/or publishing any estimate from the master file or any of the subsets, users should first determine the number of sampled respondents who contribute to the calculation of the estimate. If this number is less than 10, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. This is due to the fact that the possibility of obtaining an artificially low variance is greater with a sample size less than 10. For weighted estimates based on sample sizes of 10 or more, users should determine the coefficient of variation of the estimate and follow the guidelines described in Table 9.A.

Table 9.A: Sampling Variability Guidelines

Reliability of the estimate	CV (%)	Guidelines
Acceptable	$0.0 \leq CV \leq 16.5$	Estimates can be considered for general unrestricted release. Requires no special notation.
Marginal	$16.5 < CV \leq 33.3$	Estimates can be considered for general unrestricted release but should be accompanied by a warning cautioning subsequent users of the high sampling variability associated with the estimates. Such estimates must be identified by the **letter E** (or in some other similar fashion).
Unacceptable	$CV > 33.3$	Statistics Canada recommends not to release estimates of unacceptable quality. 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."

By definition, the CV is calculated by multiplying the standard error (equal to the square root of the estimate of the variance) by 100 and dividing the product by the estimate. Consult Chapter 11 for further information on calculating the variance.

10. Weighting

This chapter describes the weighting procedures for each subset of respondents described in Section 7.5. The longitudinal weighting process is necessarily different from that of cross-sectional weighting, for several reasons. First, longitudinal weights must represent the probability of selection of the unit of analysis at the time of sample selection. Since the longitudinal sample was selected in 1994-1995, the weights must reflect the probability of selecting the individual in Cycle 1 and not in subsequent Cycles. In addition, 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, and 2, 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, three sets of weights, WTI4LS, WTI2LF and WTI2SLF have been created. Table 10.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 10.A: Subsets of Respondents and Corresponding Sampling Weights and Flags

Subset of respondents	Sampling Weight	Flag
Longitudinal Square	W**TI**4LS	None, all records
Longitudinal Full	W**TI2**LF	W**F**I2LF
Longitudinal Full Share	W**TI2**SLF	W**FI2**SLF

The longitudinal weighting procedure is based on the weighting done for the Cycle 1. A full description of the 1994-1995 weighting methodology is provided in the public use microdata file documentation for the 1994-1995 NPHS: Health Institutions component.

10.1. Longitudinal Square Weight (WTI4LS)

The longitudinal square weight WTI4LS is mainly based on the weighting done for the Cycle 1.

Probability of Selection for Cycle 1 (1994-1995) Institutions

Notation:

M_h = number of beds in stratum *h* (based on the list of hospitals and in-patient health care institutions)

$M_{h,i}$ = number of beds in the institution *i* of stratum *h* (based on the list of hospitals and in-patient health care institutions)

n_h = number of institutions to be selected in stratum (size) *h*

The institutions were sampled from the 1994 survey frame with probability proportional to the number of beds. Consequently, in most cases, the probability of selecting an institution *i* was:

$$n_h \times \frac{M_{h,i}}{M_h}$$

When a selected institution was a head office, (refer to Section 5.1 for more details) the probability was:

$$n_h \times \frac{M_{h,i}}{M_h} \times P_{h,i,j}$$

where $P_{h,i,j}$ represents the probability that an institution *j* belonging to head office *i* is selected in stratum *h*. In the case of the largest institution belonging to *i*, $P_{h,i,j}=1$.

In the case of other institutions *j*:

$$P_{h,i,j} = \frac{M_{h,i,j}}{\sum_{j \in i'} M_{h,i,j}}$$

where *i'* represents all of the institutions belonging to head office *i,* except the largest.

Cycle 1 (1994-1995) Institutional Base Weight Calculations and Non-response adjustments at the Institutional Level

An institution's weight corresponds to the number of institutions that the sampled institution represents. The **institution's base weight** is equal to the inverse of the probability of selecting that institution. However, since there is a possibility of non-response at this level, a correction is needed to take into consideration institutions that refused to participate. In cases where interviews could not be conducted in a selected in-scope institution, an adjustment was made to the weights of the other institutions that belong to the same size stratum. This adjustment is equivalent to:

$$\frac{\textit{number of responding and non-responding institutions}}{\textit{number of responding institutions}}$$

Multiplying the initial institutional weight by this weight adjustment gives the **final Cycle 1 institutional weight**.

Cycle 1 (1994-1995) Resident Selection Probability

Notation:

$L_{h,i}$ = number of long-term residents in stratum *h*, institution *i* (obtained at time of first visit)

$r_{h,i}$ = number of residents to be selected in stratum *h*, institution *i*

Once an institution was selected, each resident in that institution had an equal probability of being selected; probability defined by:

$$\begin{cases} \dfrac{r_{h,i}}{L_{h,i}} & if\ L_{h,i} \geq r_{h,i} \\ 1 & if\ L_{h,i} < r_{h,i} \end{cases}$$

Cycle 1 (1994-1995) Residents Base Weight Calculations and Non-response adjustments at the Resident Level

To calculate the base weight applicable to residents, the final institutional weight was multiplied by the inverse of the probability of selecting a resident in the institution. Here again, because non-response is possible at this level, corrections are needed to take into consideration residents who refuse to respond (in Cycle 1). The additional correction is made to the resident base weight to take into consideration the non-response of residents:

$$\frac{\textit{sum of the weights of respondent and non-respondent residents}}{\textit{sum of the weights of respondent residents}}$$

Multiplying the base weight for residents by the non-response adjustment gives the **Longitudinal Square Weight** (**WTI4LS**).

10.2. Longitudinal Full Weight (WTI2LF)

The longitudinal full subset includes only panel members with a full response, i.e. members who have a status of "household", "deceased" or "institutionalized", at each cycle. Panel members who are excluded from this subset were therefore non-respondents, i.e. they had a status of "non-response", at some point during the first five cycles of the survey, and their weight must be redistributed to compensate for this non-response.[3]

[3] See section 7.6 for the definitions of longitudinal response pattern and full/complete response for NPHS.

The longitudinal square weight (WTI4LS) 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 WTI4LS.

The adjustments necessary in order to obtain the Cycle 5 Longitudinal Full Weight are described below.

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

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

$$\frac{\text{sum of the weights of respondent and non-respondent residents}}{\text{sum of the weights of respondent residents}}$$

This adjustment is made at the non-response classes level (in Canada). These classes are formed using CHAID (Chi-Square Automatic Interaction Detector) algorithm. This algorithm is offered with the Knowledge Seeker software (developed by ANGOSS Software International Limited).

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

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

$$\frac{\text{sum of the weights of respondent and non-respondent residents}}{\text{sum of the weights of respondent residents}}$$

This correction is made separately for each possible longitudinal response pattern (the variable LONGPAT)

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

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

$$\frac{\text{sum of the weights of respondent and non-respondent residents}}{\text{sum of the weights of respondent residents}}$$

This correction is also made separately for each possible longitudinal response pattern (the variable LONGPAT).

Adjustment 4: Adjustment for Cycle 5 (2001-2002) Non-response

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

$$\frac{\textit{sum of the weights of respondent and non-respondent residents}}{\textit{sum of the weights of respondent residents}}$$

This correction is also made separately for each possible longitudinal response pattern (the variable LONGPAT).

Adjustment 5: Post-stratification

Since the total number of people in Canada living in a health care institution is unknown (based on the institution definition in the NPHS), it is impossible to perform a post-stratification based on these totals. However, post-stratification is done using the longitudinal square (WTI4LS). Post-stratification is done in two steps: first, for each of the five regions and then for each type of institution and age-sex category.

The final longitudinal full weight **WTI2LF** is calculated by taking the longitudinal square weight (**WTI4LS**) and multiplying that value by adjustments 1 to 5.

10.3. Longitudinal Full Share Weight (WTI2SLF)

The full share subset includes the panel members who agreed to share the information provided from all interviews conducted as part of NPHS with provincial ministries of health and Health Canada but only those who had a full response in Cycles 1, 2, 3, 4 and 5. 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 in health institution.

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

$$\frac{\textit{Sum of weights for Cycles 1 to 5 responding longitudinal members in institution type / longitudinal pattern / age-sex category}}{\textit{Sum of weights for Cycles 1 to 5 responding longitudinal members who agreed to share, in an institution type / longitudinal pattern / age-sex category}}$$

Note that a few of the original longitudinal response patterns were collapsed in order to produce more stable adjustments. The grouping was done for a few institution type/age-sex categories that had few observations in some of the longitudinal patterns. In each case, the problematic response pattern was grouped with another longitudinal pattern in the same institution type/age-sex category, so that the sum of the weights would still give the correct population counts. The grouping was also done, as much as possible, in

the same Cycle 5 response status. For example, if the problematic case had a response status "dead" in Cycle 5, the case has been grouped with another longitudinal response pattern that had also a response status "dead" in cycle 5. For some cases, it was not possible to do such a grouping therefore, instead of grouping response pattern, age-sex categories were grouped.

The final longitudinal full sharing weight, **WTI2SLF**, is obtained by multiplying the longitudinal full weight, WTI2LF, by this adjustment. In general, since this adjustment is done with respect to the post-stratification classes, no additional post-stratification is necessary.

11. Calculation of Variance

The method used to calculate the variance of estimates in Cycle 5 is the same as the one used in Cycle 3 and 4 (different from Cycle 2 method). Since Cycle 3, the bootstrap method has been used. This method is used for the NPHS Households component and is explained in this section.

A variance calculation program, developed for SAS or SPSS, is provided with the data file. It can be used to obtain specific estimates of variance for such statistics as totals and ratios and for more complex analyses, such as regressions. A user guide is provided with the program.

11.1. Bootstrap Method

The sampling designs for health surveys are complex. Since the variance for such designs cannot be calculated with simple formulas, a re-sampling method is necessary to calculate the variance. The bootstrap method consists of sub-sampling the initial sample. Within each stratum, a simple random sample (SRS) is selected, with replacement, from n-1 clusters within the n clusters of the stratum. This creates B new samples (or repetitions). The same estimate is then calculated for each of the B samples, which gives B different estimates. To obtain each of the B estimates, a specific weight for each sample is necessary. In each SRS sample, the weight is then recalculated for each record in the stratum. These B weights, the bootstrap weights, have been produced and are available with the data (for the longitudinal full subset only, due to the small difference in the number of records in the longitudinal full and square subsets).

In summary, the bootstrap method consists of:

A) Calculating an estimate (total, ratio, etc.) using the final weights included in the data file. This estimate is the point estimate.

B) Calculating the same estimate, this time using each of the B bootstrap weights contained in the bootstrap files. B estimates (total, ratio, etc) are then obtained.

C) Finally, calculating the variance of the B estimates. This variance is the estimate of the variance of the point estimate calculated in A.

The rules for confidentiality and release guidelines also apply to the variance estimates obtained through the bootstrap method.

11.2. Estimating Variance with the BOOTVAR.SAS (.SPS) Program

BOOTVAR.SAS (.SPS) is the program used to estimate the variance. This program comes with the data file. It is used to estimate the variance for totals, ratios, differences between ratios, parameters of linear and logistic regressions, and general linear models. The user must ensure the references to the file names are consistent when using the program. For more information on how to use BOOTVAR.SAS (.SPS), consult the user guide provided with the program.

12. Using the Longitudinal Master Files

12.1. Use of Longitudinal Weights

In the first three cycles, files made up of subsets of the 2,287 longitudinal panel members were created. For cycle 5, as in cycle 4, only one file has been produced. The Cycle 5 master file contains three subsets of respondents to which corresponds a set of weights. Flags were created to identify records that are part of a particular subset.

Table 12A: Subsets of respondents, weight variable and corresponding flag

Subset of respondents	Weight	Flag variable
Longitudinal square	W**T**I4LS Master weight – Longitudinal	None, all records
Longitudinal full	W**T**I2LF Master weight – Longitudinal full response	W**F**I2LF
Longitudinal full share	W**T**I2SLF Share weight – Longitudinal full response	W**F**I2SLF

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 WTI4LS is called the Master weight – Longitudinal, also known as the "square weight" and applies to the "Square" subset of respondents which includes all 2,287 members that make up the original longitudinal sample. This weight is to be used if users wish to do specialized studies on non-response bias.

Weight WTI2LF is called the Master weight – Longitudinal full response, also known as the "Longitudinal Full" weight and applies to the 2,131 records that are included in the "Full" subset of respondents.

Weight WTI2SLF is called the Share weight – Longitudinal full response, also known as the "Longitudinal Full Share" weight and applies to the 2,067 records that are included in the "Full Share" subset of respondents.

NOTE: for the NPHS – Health Institutions component, the definition of "full" includes partially complete responses as well as fully complete responses.

12.2. Ensuring the Reliability of Estimates with the Use of Bootstrap Weights

Bootstrap weights are necessary for variance estimation. Information on the bootstrap method for variance estimation can be found in Chapter 11. Bootstrap weights are available for the "Full" subsets (Master and Share) only.

Due to the complex sample design, users should use the Bootvar program for variance calculation. The standard variance output from other statistical packages such as SAS and SPSS may grossly underestimate the variance of an estimate for this survey. **It is the responsibility of the user to ensure the quality/reliability of the estimates that they are producing by following the guidelines laid out in Chapter 9 and correctly calculating the variance for all estimates.** Failure to do so could lead to some misinterpretation of results and jeopardize the quality of the research work. Some statistical software are capable of including the stratum and cluster information as input when performing analytical processing, which does provide a variance estimate much closer to the true variance estimate, but these packages fail to account for the various weighting adjustments, which in some cases can impact the variance estimates considerably.

12.3. Variable Naming Convention

In 1996-1997, NPHS adopted a variable naming convention, which 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 applied: restrict variable names to a maximum of 8 characters for ease of use by analytical software products; identify the survey cycle (1994-1995, 1996-1997, 1998-1999, 2000-2001, 2002-2003) in the name; and allow conceptually identical variables to be easily identifiable over survey cycles. For example, conceptually identical data on smoking were collected in 1994-1995 and 1996-1997. The variable names about smoking should only differ in the year position. This convention will be followed throughout the longitudinal survey, and will be adopted by all NPHS components: the household survey, the institutional survey, the North component survey, and supplements.

12.3.1. Variable Name Component Structure

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

Positions 1-2:	Variable / Questionnaire section name
Position 3:	Survey type / component
Position 4:	Year / cycle in which the variable appears
Position 5:	Variable type (i.e., questionnaire, coded, derived, etc.)
Positions 6-8:	Variable number / name from questionnaire

For example, the name of the variable DHI2DAGE means:

DH:	found in the Demographic and Household content section of the questionnaire
I:	questions that are on the Institutions survey
2:	appeared in Cycle 5 (2002-2003)
D:	derived variable
AGE:	variable name

12.3.2. Positions 1-2: Variable / Questionnaire Section Name

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

AL	Alcohol	HW	Height and Weight
AM	Administration of the survey	IN	Income
CC	Chronic conditions	IP	Institutions Policies
DG	Drug use	RA	Restriction of activities
DH	Demographics and household	SD	Socio-demographics
ED	Education	SM	Smoking
FL	Balance and falling	SP	Sample identifiers (methodology)
GH	General health	SS	Social support
HC	Health care utilization	WT	Weights
HS	Health status		

12.3.3. Positions 3: Survey Type / Component

A Asthma supplement
B Province-specific buy-in content - children's questions
C Household Core questions that will be repeated in each cycle
I Institutions component
K Longitudinal children's questions
N North (Yukon / NWT) component
P Province-specific buy-in content - adult questions
S National supplement (Health Promotion Survey)
_ Cycle specific questions, not repeated in every cycle (stress in 1994-1995, access to services in 1996-1997)
3 Survey administration variables at the household level in the household component (H03)
5 Survey administration variables for the General file of the household component (H05)
6 Survey administration variables for the Health file of the household component (H06)

12.3.4. Position 4: Year / Cycle

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)

12.3.5. Position 5: Variable Type

_	Collected variable	A variable that appeared directly on the questionnaire
C	Coded variable	A variable coded from one or more collected variables (e.g., SIC, Standard Industrial Classification code)
D	Derived variable	A variable calculated from one or more collected or coded variables, usually calculated during head office processing (e.g., health status index)
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)
G	Grouped variable	Collected, coded, suppressed or derived variables collapsed into groups (e.g., age groups)
L	Longitudinal derived variable	A variable calculated using variables from two or more survey cycles

12.3.6. Positions 6-8: Variable Name

In general, the last three positions follow the naming on the questionnaire. Numbers are used where possible: Q1 becomes 1. "Mark-all" questions use letters for each possible answer category: Q1 (mark all that applies) 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 DHI2DAGE in 2002-2003. 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. An 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.

12.4. Access to Master File Data

12.4.1. Microdata Files

Confidentiality concerns preclude general dissemination of longitudinal NPHS – Health Institutions data in public use microdata file (PUMF) format. However, on-site access to the NPHS – Health Institutions master microdata files is possible at

Statistics Canada's Research Data Centres (RDCs). These centres, established in collaboration with the Social Sciences and Humanities Research Council (SSHRC), are situated in secure physical locations at participating universities. They operate as extensions of Statistics Canada offices, with a full-time Statistics Canada employee at each centre, and researchers conduct their work under the terms of the *Statistics Act*, as would any other Statistics Canada employee. More information is available at the Research Data Centres Program web page: http://www.statcan.ca/english/rdc/index.htm.

PUMFs are available for each of the first two cycles of the NPHS – Health Institutions component, providing 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, and 2 NPHS – Health Institutions cross-sectional 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.

12.4.2. Tabulations

To access the survey Master file, one approach for any client is the production of custom tabulations done by the Client Custom Services staff in Health Statistics Division. This service allows users who do not possess knowledge of tabulation software products to get custom results. The results are screened for confidentiality and reliability concerns before release. There is a charge for this service. You can access this service by writing to hd-ds@statcan.ca , calling 1-613-951-1746 or by fax 1-613-951-0792.