1998-1999 NATIONAL POPULATION HEALTH SURVEY
HEALTH INSTITUTIONS COMPONENT
TABLE OF CONTENTS

1. Introduction

The National Population Health Survey (NPHS): Health Institutions component is the first national longitudinal survey of residents of Canadian health care facilities. The third cycle of data collection began in February 1999 with the second follow-up of the longitudinal respondents interviewed in 1994-1995 for the first cycle of the survey. Unlike Cycle 2, no cross-sectional top-up was selected for Cycle 3. Also new for Cycle 3, the collection of data from persons in the institutions sample who moved to households was done by the Health Institutions component, rather than the Households component as was done in Cycle 2. Questions for household residents were separated from those of the institution residents with over 90% of the questions identical.

This manual has been produced to facilitate the use of the Cycle 3 (1998-1999) master file of the NPHS: Health Institutions component.

Any questions about the data set or its use should be directed to:

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

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

For remote access to NPHS master files call: 1-613-951-1653
Colette Koeune, NPHS, Health Statistics Division Internet: nphs-ensp@statcan.ca
Fax: 1-613-951-4198

For survey content support call:
Mario Bédard, NPHS, Health Statistics Division 1-613-951-8933
Internet: mario.bedard@statcan.ca
Fax: 1-613-951-4198

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. The questionnaire included components 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 3 (1998-1999)

The following is a list of the items that were modified, added or dropped:

- The questions asked to respondents depend on which sample they belong. If the respondent lives in an institution, he will answer questions from sections B to P. If the respondent was coming from sample I (Institutions) but went back home, he will answer questions from sections CC to QQ.

- *Regional Office List of Institutions* (Form 1) and *Selected Information Form* (Form 2) were not needed since there was no cross-sectional top-up. The final status codes and contacts with the institution are on the cover of the *Institution Control Form*.

- The *List of Long-term Residents* (Form 99), was no longer needed as it summarized the admission data for residents who were admitted before and after the end of Cycle 1 collection, in order to select a top-up sample, however no top-up was done in Cycle 3. Interviewers did receive a spreadsheet with additional tracing and contact information for each institutional resident and each household respondent.

- The content of Form 3 – *Status of the Longitudinal Respondent*, was incorporated into the Respondent Questionnaire (Section A) to simplify collection of both the institution longitudinal and core (household) longitudinal respondent information.

- The confirmation of sex, date of birth and Provincial health number were dropped since they had been confirmed in Cycles 1 and 2.

- Height and Weight were dropped, since they were used to calculate the Body Mass Index (BMI), which we do not calculate for people aged over 65.

- *Consent to Interview* (given by next-of-kin) was incorporated into the Respondent Questionnaire in Sections B, N and QQ. Where written consent was needed by the institution, a general statement for the release of information was added to the *Consent for Institution to Release Information Form*. This form was changed to incorporate wording from the core, and to drop the reference to the collection of height and weight data.

- The date of status interview was dropped by mistake. A variable with a fixed date was created during processing to replace this variable.

- In the *Institution Control Form* (previously form 5), the form number was dropped, the address fields were added, the interviewer number was dropped (it duplicated assignment number), and the sequence number (pre-printed on forms) was added to link the control forms and the questionnaires.

- The *Resident Questionnaire* (Form 6) was re-named to *Respondent Questionnaire* (the form number was dropped), for Cycle 3 since it was used for both residents and people in households.

- "Sample type = C or I" was added to the questionnaire label in order to distinguish core (C) and institution (I) samples.

- The location and sequence number questions were added to the cover page of the *Respondent Questionnaire* to aid in the linking of questionnaires and control forms.

- A final status code 077 was added for core sample movers back to household. The

household component interviewers will trace these cases in Q5.

- The introduction wording changed to correspond to the introduction of the household survey.

- The question wording changed from "Is… still a long-term resident of this facility?" to "Does …still live at (read address on label)?" in order to incorporate the wording for household respondents.

- The category "private home" became " private household" to further emphasize that the move be out of institutional care and into a household setting. "Another residential care facility" became "Residential care facility" in order to incorporate household respondents.

- In *Status of Longitudinal Respondent* (Section A), the questions on the city and province of death were added to help facilitate the match with the mortality files. A question on the telephone number of the institution was added to facilitate tracing and a major skip for type of sample and location was added to incorporate the core sample movers and the institution sample household residents.

- The *Section for Selected Resident Information* (Sections C and CC) was changed to *Selected Respondent* to be consistent with the terminology for household residents. Respondents living in households do not receive questions C4, C5, C6 and C7.

- There were no changes to the section *General Health* (Sections D and DD), nor Health Status (Sections E and EE).

- Cancer was added back as a chronic condition (it had been included in Cycle 1), to the *Chronic Conditions* (Sections F and FF).

- For the section *Restriction of Activities* (Sections G and GG), the question RA-Q1 (GG1) for people living in households takes the households core questionnaire wording. The question RA-Q7 (GG8) uses the institution question, but replaces "institution" with "home". The question RA-Q5 on the second condition or health problem was separated into two questions: G5/GG5 are Yes/No questions, then G6/GG6 is the long answer for the second condition.

- *Balance* (Sections H and HH): in the question FL_Q4 (H4/HH4), "commotion cérébrale" was added to the French questionnaire.

- *Smoking* (Sections I and II): no change.

- *Alcohol* (Sections J and JJ): the question AL-Q2 and AL-Q3 changed places. Skip pattern changes.

- *Social Support* (Section K): some "None" categories were replaced with interviewer instruction "If None enter 00". Skip SS_Q11 (K12) was changed to allow those who leave the institution for outings less than once a month to receive the question on the type of outings, SS_Q12 (K13). Two new categories were added: "Go for a drive" and "Go out for lunch or dinner". The question on flexibility of schedule was added back in from Cycle 1.

- *Socio-demographic Characteristics* (Section L and LL): “Unemployment insurance” was changed to “Employment insurance”, and country of birth, ethnicity, race, language and education questions were dropped.

- *Contact Information* (Section M and MM): the province code was added to the contact addresses and “employee of facility” was added to the list of contact persons.

- *Drug Use* (Section O and OO): the name of the staff member was developped and an interviewer instruction was added to “Ask the person to look at the bottle, tube or box.”

- In *Health Care Utilization* (Section P and PP), the persons in households are not asked about transfers to acute care facilities.

5. Sample Design

The longitudinal sample of the 1998-1999 NPHS Health Institutions component consists of all longitudinal respondents chosen in cycle 1. This sample consists of 2,287 persons living in a health care institution in 1994-1995.

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. A list of in-scope long-term health care institutions was created and then stratified by geographic region, type of institution and size of institution. Institutions considered wer ethe long term institutions where health care were the main services offered. Institutions must have at least 4 beds and residents cannot be autonomous. A sample of institutions was chosen from this list and then a sample of residents was chosen whithin each selected institution. 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.

5.1 1994-1995 Sample Stratification and Allocation

The cross-sectional 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.

The list of health institutions was initially stratified by geographic region (geographic stratum) and subsequently by the type of institution (characteristic stratum) and number of beds (size stratum).

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. 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.

5.2 1994-1995 Sample Selection

In cycle 1, the number of institutions selected from a size stratum depended on the amount of sample allocated to the stratum 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.

5.3 1998-1999 Sample

The size of the sample is gradually diminishing due to the erosion of the panel since 1994. The NPHS defines erosion as non-response, refusals and persons who could not be located. The decline in the size of the sample has been only slight and should not increase the variance in the estimations. However, it is important to note that while deceased persons were excluded from 1998-1999 collection, they remain valid for the longitudinal analyses and are included in the master file. Deceased persons are considered as complete responses for longitudinal analysis. Table 1 shows the size of the longitudinal sample of the 1998-1999 NPHS Health Institutions component.

Table 1		
Size of the longitudinal sample		
1994 longitudinal sample	1998 full longitudinal file: *complete response in 1994, 1996 and 1998*	Deceased persons (on 2,178)
2,287	2,178	1,250

Since no longitudinal unit was selected in cycle 3, the population covered by the longitudinal sample represents the 1994-1995 population.

The status of the members of the longitudinal survey may have changed since 1994. The various statuses are summarized in Table 2.

Table 2

Status of members of the longitudinal panel in 1998	Action taken
Still resides in cycle 1 institution	Longitudinal panel members who still reside in their cycle 1 institution were interviewed.
Moved to another institution	Longitudinal panel members who moved to another institution were interviewed (unless they had moved to a correctional institution).
Moved to a private household	Longitudinal panel members who moved to a private household were interviewed.
Deceased	Deceased members of the longitudinal panel were excluded from the collection but are part of the longitudinal file of the health care institutions component.

The other categories for members of the longitudinal panel are as follows: moved to the Northwest Territories or the Yukon; moved to an Indian reserve or to a Canadian Forces Base; and moved outside of Canada (temporarily or permanently). For the health care institutions component, none of these situations occurred in the third cycle. These types of changes are normally noted among the members of the population living in private households rather than among persons living in an institution.

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 57% of the interviews (of the proxy interviews, 71% 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 February until May 1999. 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 program 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. 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 all the selected institutions participating in the survey.

7. Data Processing

7.1 Data Capture and Editing

The respondent questionnaire was captured manually in the Regional Offices using DC2. The Programmes written for the data capture of the questionnaires prevented out-of-range values from being entered.

The Institution Control Form (ICF) was captured manually. The information was 100% verified because of the small number of records.

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.

After data capture, questionnaire data flows were verified and consistency edits between certain fields were performed. With the exception of the Health Utility Index (HUI), no imputation was performed (see Section 8.3).

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). Drugs and medications were coded using a revised version of the Canadian Anatomical Therapeutic Chemical Classification System (ATC) developed by Health Canada.

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 these variables were created can be found in Appendix D.

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.

8.1.1 Cycle 1 Response Rate (1994-1995)

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

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

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

$$= 95.5\%$$

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:

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

$$= \frac{2287}{2444} \times 100$$

$$= 93.6\%$$

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

8.1.2 Cycle 2 Response Rate (1996-1997)

All of the institutions that participated in the first cycle of the survey and that were still in operation and all institutions newly covered agreed to have their residents surveyed. The institutions response rate is calculated as follows:

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

$$= \frac{314}{314} \times 100$$

$$= 100\%$$

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:

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

$$= \frac{2193}{2287} \times 100$$

$$= 95.9\%$$

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

8.1.3 Cycle 3 Response Rate (1998-1999)

All of the institutions that participated in the first cycle of the survey and that were still in operation and all institutions newly covered by the survey agreed to have their residents surveyed. The institutions response rate is calculated as follows:

$$\frac{\textit{Number of institutions selected that agreed to participate}}{\textit{Total number of institutions where panle people resided}} \quad x\ 100$$

$$= \frac{352}{352} \times 100$$

$$= 100\%$$

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:

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

$$= \frac{2251}{2287} \times 100$$

$$= 98.4\%$$

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

8.2 Survey Errors

8.2.1 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:

$$\left(\frac{0,0094}{0,104}\right) x\ 100\ \% = 9,04\ \%$$

Section 11 of this documentation 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.

8.2.2 Non-sampling Errors

Errors not related to sampling may occur at almost every stage of a survey. Interviewers may misunderstand instructions, respondents may make errors answering the questions, answers may be incorrectly entered in the computer or errors may in 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 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 were 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 (refere to section 10 for more information on weighting).

8.3 Imputation

Imputation was used to derive the missing values for one variable in the NPHS Health Institutions component. The variable HSI8DHSI denotes the resident's Health Status Index (HSI) 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 HSI rating, which can range from 0 to 1 (with the possibility of having negative values), are calculated based on responses to a series of questions on health status.[1] However, this overall rating cannot be calculated if one or more of the answers are missing, a situation that occurs for about 5% of respondents. It was decided to use imputation for the missing values in order to calculate the HSI in the 1998-1999 cycle of the health care institutions component. We therefore used the **hot deck method** to impute values for the missing elements in order to be able to calculate the overall HSI for the individuals concerned. It should be noted that the method used was the same one used in 1994 and 1996.

The HSI 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 HSI rating. Imputation was at the level of the eight partial ratings rather than the questions. After imputation, the program for calculating the derived HSI 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 three 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

[1] For more information on the calculation of the HSI, see Appendix D: Derived variables.

deduce the partial rating with certainty. We therefore attributed a partial rating based on this partial information 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. Regression models were used to analyse the records containing valid values for each of the eight partial ratings in order to determine what other variables of the questionnaire might explain the value of the partial ratings and might therefore be used to predict the value of the missing partial ratings. We used these explanatory variables as matching variables to identify the donor records for imputation in the records missing values for certain partial ratings.

- Lastly, in some cases, the imputed value of a partial rating did not match the partial information appearing on the receiving record. For example, partial information on a record might indicate that the partial rating for the hearing element should be between 2 and 3. When the imputed value was outside this range, the value closest to the imputed value falling within the acceptable range was used for imputation. In the example given, if the imputed value was higher than 3, it was changed to be equal to 3, and if it was below 2, then it was changed to equal 2.

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 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 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

SMI8_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

SMI8_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 by the denominator estimate.

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. We multiply the value reported in question SMI8_3 by the weight for the record (WTI8LF), then we sum this value over all records with a response of "daily" to SMI8_1.

To obtain a weighted average expressed in the form $\hat{Y} / \hat{X}$, we calculate the numerator ($\hat{Y}$) 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 SMI8_1 is "daily";

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

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. 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 which 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 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 estimates from this file, users should first determine the number of respondents who contributed to the calculation of the estimate. If this number is less than **30**, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. For weighted estimates based on sample sizes of 30 or more, users should determine the coefficient of variation for the estimate using the SAS or SPSS program of variance estimation provided (see section 11.1) and follow the guidelines below.

Sampling Variability Guidelines

Reliability of the estimate	CV (%)	Guidelines
1. Acceptable	0.0 to 16.5	Estimates can be considered for general unrestricted release. Requires no special notation.
2. Marginal	16.6 to 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 M (or in some other similar fashion).
3. Unacceptable	Greater than 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 U (or in some other fashion) and the following warning should accompany the estimates: "The user is advised that . . . (specify the data) . . . do not meet Statistics Canada's quality standards for this statistical program. Conclusions based on these data will be unreliable and most likely invalid. These data and any consequent findings should not be published. If the user chooses to publish these data or findings, then this disclaimer must be published with the data."

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 section 11 for further information on calculating the variance.

10. Weighting

Unlike cross-sectional weighting, longitudinal weighting considers the probability of selection of the unit for analysis at the time of the selection of the sample. The weights attributed to the respondent units of the NPHS Health Institutions component are based on the probability of having selected the unit at the time of the sample selection in 1994.

Weighting for cycle 3 has been largely based on weighting for Cycle 1. A full description of the 1994 weighting methodology is provided in the public use microdata file documentation for the 1994-1995 NPHS: Health Institutions component. Some adjustments have been made in order to correct for non-response observed in cycles 2 and 3.

The section below provides a brief description of the 1994 weighting method, which remains valid for cycle 3 longitudinal weighting, along with a description of the specific cycle 3 adjustments.

10.1 Probability of Selection for 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); and
- 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:

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

The probability of selecting a head office (refere to section 5.3 for more details) 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 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.

10.1.1 1994-1995 Institutional Weight Calculations & 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**.

10.2 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:

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

10.2.1 1994-1995 Base Weight for Residents and Correction for Non-response at the Resident Level

To calculate the base weight applicable to residents, we multiply the final institutional weight 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 institution final weight by the base weight for residents corrected for non-response gives the **resident weight for cycle1**.

10.2.2 Correction for Non-response in 1996-1997

Non-response is also possible in cycle 2. This cycle non-respondents are not part of the cycle 3 full file. Their weights must then be redistributed to cycle 2 respondents. The cycle 1 weights of cycle 2 respondent are multiplied by:

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

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).

We then obtain **resident weight in cycle 2**.

10.2.3 Correction for Non-response in 1998-1999

A correction is also made to take into account non-response in cycle 3. The residents weights (resident weights in cycle 2) is multiplied by:

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

This correction is made separately for each possible longitudinal response category (the variable LONGPAT, i.e. the type of institution for each of the cycles).

10.2.4 Post-stratification Adjustment

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 total weights obtained in cycle 1. Post-stratification is done in two steps. First, for each of the five regions, then for each type of institution and age-sexe category.

The combination of these adjustments gives the **final resident weight (WTI8LF)**.

11. Calculation of Variance

The method used to calculate the variance of estimates in cycle 3 is different from that used in cycle 2. In cycle 3, the bootstrap method was used. This method was used for the NPHS Households component and is explained in this section.

A variance calculation program, developed for the SAS system, 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 resampling method is necessary to calculate the variance.

The bootstrap method consists of subsampling 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.

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

11.2 Estimating Variance with the BOOTVAR.SAS Program

BOOTVAR.SAS 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, consult the user guide provided with the program.

12. File Use

12.1 Variable Naming Convention

In 1996-97, 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...) 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 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.

12.1.1 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 DHI6GAGE means:

DH:	found in the Demographic and Household content section of the questionnaire;
I:	questions that are on the Institutions survey;
6:	appeared in 1996-1997 cycle;
G:	grouped variable; and
AGE:	variable name.

12.1.2 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		

12.1.3 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)

12.1.4 Position 4: Year / Cycle

4 1994 - 1995
6 1996 - 1997
8 1998 - 1999
0 2000 - 2001
2 2002 - 2003
A 2004 - 2005
B 2006 - 2007
C 2008 - 2009
D 2010 - 2011
E 2012 - 2013

12.1.5 Position 5: Variable Type

_	Collected variable	A variable that appeared directly on the questionnaire
C	Coded variable	A variable coded from one or more collected variables (e.g., SIC, Standard Industrial Classification code)
D	Cross-sectional derived variable	A variable calculated from one or more collected or coded variables, usually calculated during head office processing (e.g., health status index)
F	Flag variable	A variable calculated from one or more collected variables (like a derived variable), but usually calculated by the 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

12.1.6 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 DHI6GAGE in 1996-97. 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.

12.2 Access to Master Files Data

There are three ways of accessing the survey master files. The first way is to use the remote access to the survey master file. The user can be supplied with a 'dummy' test master file and a corresponding record layout. With this, the user can spend time developing his or her own set of analytical computer programs using the test file to confirm that the routines are functioning correctly. At that point, the code for the custom tabulations is then sent via the e-mail to nphs-ensp@statcan.ca. The code will be moved into Statistics Canada's internal secured network and processed against the appropriate master file of NPHS Health Institutions component data. Results are screened for confidentiality and reliability concerns and, once these have been addressed, the output is returned to the client. There is no charge for this service.

A second approach for any client is the production of custom tabulations done by the Client Custom Services staff in Health Statistics Division. This service allows users who do not possess knowledge of tabulation software products to have access to the master file for the preparation of their own custom calculations. As with remote access, the results are screened for confidentiality and reliability concerns before release. Unlike remote access, there is a charge for this service.

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