NATIONAL POPULATION HEALTH SURVEY
HEALTH INSTITUTIONS 1996-1997
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 second cycle of data collection began in the fall of 1996 with the follow-up of the residents interviewed in 1994-95 for the first cycle of the survey. Because of the high attrition rate due to deaths in the original sample, an additional sample was selected in the original 1994 facilities and new facilities were selected. This ensures that representative cross-sectional estimates could be produced in addition to the longitudinal data.

The initial follow-up of residents was conducted in October and November 1996. The interviews were conducted between December 1996 and April of 1997 with additional follow-up of residents who moved continuing until June 1997. This manual has been produced to facilitate the use of the 1996-97 cross-sectional microdata file of the survey results.

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

For survey content support call:
Mario Bédard, NPHS, Health Statistics Division 1-613-951-8933
Internet: bedardm@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 Institutional 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 relating 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 1996-97

There was little change to the questionnaire between 1994-95 and 1996-97. The following is a list of the items that were modified, added or dropped:

- Demographic information was printed on a label and affixed to the questionnaire to be confirmed in the course of the interview.

- Institution policy questions regarding husbands and wives sharing a room and pet visits were dropped. Questions on the type of activities organized for residents, the organization of activities for residents' families and the completion of living wills were

added in 1996-97.

- A question assessing the resident's agility was added to the Health Status Section.

- The list of chronic conditions was revised so that it was more comparable to the household list and so that it better reflected the conditions that were present in the institutional population based on the results of the 1994-95 survey.

- A question on whether respondents in the institution had a telephone in their room was added in cycle 2.

- The Restrictions of Activities section was modified to make it comparable with the household questionnaire. In cycle 1, four limiting conditions were captured and coded; in cycle 2, this was reduced to two.

- There were a few revisions in the Social Support section; questions were added to determine if the resident took part in any one-to-one activities with a staff member and in what activities the resident had taken part while outside the facility. They were also asked about the frequency of their telephone contacts with friends or relatives. The question regarding the flexibility of the daily schedule was dropped.

- The categories in the education question were changed to make them more comparable with the derived variable categories from the household survey.

- In addition to obtaining the drug and utilization information from the institution, we also collected the residents' height and weight (these do not appear on the public use microdata file) and whether they had been transferred to a short-term care facility for a period of fewer than 21 days within the past 12 months.

5. Sample Design

The target population of the institutional survey included all long-term residents of health institutions from all provinces, excluding the territories, Indian Reserves and Canadian Forces bases. A list frame of facilities with long-term residents was created and stratified by geographic region, type and size of facility. A sample of institutions and a subsequent sample of residents within these institutions were selected. Those institutions that are not health-related are not part of the institutional frame. This includes correctional institutions, jails, young offenders' facilities, as well as children's group homes (orphanages) and religious institutions.

5.1 1996-97 Sample Design

In the first cycle of the NPHS: Health institutions, the sample was created by first selecting institutions and then residents within these institutions. For the second cycle, a distinction is made between the sample selected for longitudinal purposes and the sample selected for cross-sectional purposes.

The cross-sectional sample consists of following-up the longitudinal respondents and selecting a supplementary cross-sectional sample. The longitudinal sample for 1996-97 consists of all longitudinal respondents chosen in cycle 1 who had responded to the questionnaire in 1994-95. For health institutions, the attrition of the sample is expected to be much higher than for people living in private households (see Section 5.1.1 below). The sample attrition of selected residents of health institutions is mainly due to the death of longitudinal panel members. Since the decrease in sample size is substantial, it can lead to large increases in the variance estimates if no additional sample is selected. Moreover, in order to maintain cross-sectional representativity of our sample, a sample of newly admitted residents in cycle 1 institutions and a top-up sample of new institutions must be drawn. The follow-up procedures for the longitudinal panel are first described below.

5.1.1 Follow-up of the Entire Longitudinal Panel

There were two phases of data collection in cycle 2. In phase 1, the interviewers returned to the health care institutions selected in cycle 1 in order to determine if selected longitudinal panel members still resided in their facility or had moved or died. The administrators or contacts from all the health care facilities selected in cycle 1 were contacted in October and November 1996. They were asked to complete a paper questionnaire on the status of the longitudinal panel members (to obtain the date of death for deceased panel members or tracing information in the cases of movers) and compile a list of all long-term residents of the facility with their dates of admission. The interviewer then completed a Selection Information Form giving the number of long-term residents admitted before April 1, 1995, the

number admitted on or after April 1, 1995, the number of longitudinal panel members who were still in the facility, and the number who were no longer in the facility. This cut-off date was chosen because the last day of data collection in the first cycle was March 31, 1995. Newly admitted residents could then be identified. Results from phase 1 showed that over one-third of the original panel members were no longer residing in their cycle 1 institutions. Of these, almost three-quarters of cases were deceased panel members and the balance movers (in fact it was found later during tracing that some of these movers were also panel members who had died). All the information was sent back to Head Office so that the additional sample for the 1996-97 cross-sectional file could be selected.

The 1996-97 cross-sectional sample comprises cycle 1 institutions still in operation as well as new top-up institutions selected for cycle 2. The following Table 1 gives the rules to identify whether longitudinal panel members were part of the 1996-97 cross-sectional file.

Table 1

Status of Longitudinal Panel Member	Action Taken
Still resided in cycle 1 institution	Longitudinal panel members who still resided in their cycle 1 institutions were interviewed and are part of the cross-sectional file.
Moved to another institution	Longitudinal panel members who moved to another institution were interviewed (if they moved to a correctional facility no attempt was made to interview) but are not part of the cross-sectional file. Their data appear only on the longitudinal institutional file.
Moved to a private household	Longitudinal panel members who moved to a private household were interviewed by the Household component of NPHS. They are not part of the cross-sectional file. Their data appear on the longitudinal institutional file.
Deceased	For longitudinal panel members identified to be deceased at the time of cycle 2 interview, the death was confirmed against the Canadian Mortality Database. Longitudinal panel members who have died are part of the longitudinal institutional file, but do not appear on the cross-sectional file.

Other possibilities for the status of the longitudinal panel member are the following: moved to the Northwest Territories or the Yukon, moved to an Indian reserve or Canadian Forces base, and moved out of Canada (either temporarily or permanently). None of these situations occurred in cycle 2 for the institutional panel. Such movements are generally observed for the population of household residents rather than for institutionalized persons.

5.2 Design of the 1996 Frame

The 1996 sample frame was generated from the 1995 list of residential care facilities collected by the Canadian Institute for Health Information (CIHI) and the 1993 list of hospitals maintained by the Health Statistics Division (HSD) of Statistics Canada. Provincial Ministries of Health verified and updated these lists to ensure their accuracy. The institutions were classified by the dominant type of care provided and only those providing long-term care were retained[1]. From the residential care facilities list, institutions providing long-term care for aged people, emotionally disturbed children, developmentally delayed, physically and psychiatrically disabled people were retained. Facilities from the hospital list included general hospitals with long-term units, extended/chronic care or rehabilitation facilities, and speciality hospitals with long-term units, such as pediatric and psychiatric hospitals. The number of long-term beds was known for each institution.

The sample population was restricted to those facilities with at least four beds that provided long-term care to residents with health problems. Facilities with fewer than four beds were not included on the CIHI list of residential care facilities. Each residential care facility is also classified by the level of care provided at the facility. When the level of care is lower than Type 1 then the facility is deemed to contain self-sufficient residents and is therefore excluded from the frame. Type 1 care is defined as care that is required by a person who is ambulant and/or independently mobile, who has decreased physical and/or mental faculties, and who requires primarily supervision and/or assistance with activities of daily living. Note that for the province of Québec the level of care is not provided. Following discussions with members of the Québec Health Ministry, it was determined that the residential care facilities list generally do not contain self-sufficient institutions in this province.

[1] Institutions that exclusively provided short-term care, such as drug rehabilitation centres, were excluded because the household component of the NPHS covers short-term institutional residents.

5.2.1 Identification of new institutions on 1996 Frame

The creation of the 1996 frame is based upon the same criteria and stratification (see Section 5.3 below) used for the 1994 frame. The 1994 and 1996 list frames of provincial long-term health care facilities were matched to identify new institutions. A new sample was drawn from these facilities (see Section 5.4 for sample selection). One should note that the definition of a new institution does not necessarily equate with newly established institutions since the last survey. Rather, it represents all institutions on the 1996 frame that were not found on the 1994 frame. Some institutions are now in scope on the basis of a higher level of care in 1996 (usually passing from level 3 to level 4 (or Type 1)).

5.3 Stratification and Allocation

The sample allocation for the second cycle depends significantly on that of the first cycle. In the following section the stratification and the sample allocation of the first cycle are described.

5.3.1 1994-95 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.

5.3.2 1996-97 Stratification and Allocation

As mentioned in Section 5.2, the same stratification (geographic region, type of institution, number of beds) as that used in cycle 1 was kept to classify the new health care facilities. Based on the information provided by the follow-up of the longitudinal panel members, we decided how many extra residents needed to be interviewed in each cycle 1 institution, and an additional sample of residents was randomly selected. The sample size of the top-up sample in each cycle 1 institution depends upon the sample attrition for that institution, the number of long-term residents in 1996, and the corresponding 1994 sampling fraction.

For each new institution, the sample size was first set to the number of residents sampled per institution in 1994 for that stratum. The sample size was then adjusted to account for non-response at the person level. That adjustment was based on the cycle 1 individual response rate, which was 93.6%. Table 2 below gives the 1996-97 final sample allocation by type of institution.

Table 2

Sample Size by Type of Institution

Aged	Cognitive	Other Rehabilitative	Total
1,816	339	238	2,393

5.4 Sample Selection

5.4.1 1994-95 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.

5.4.2 1996-97 Sample Selection for cycle 1 institutions

From the list of long-term residents in cycle 2, two sub-lists were compiled for each selected institution. Both lists were based on the date of admission. The first list gives all residents who were present at the time of the cycle 1 survey (i.e. admitted before April 1, 1995). Longitudinal respondents still residing in the institution had been removed previously from that list. The second list gives all newly admitted residents (i.e., admitted on or after April 1, 1995). A systematic sample of residents was chosen from the list of new residents. For some institutions another systematic sample of residents was selected among the list of residents present at the time of cycle 1 survey. The top-up sample selected by Head Office was then sent back to regional offices for interviewing.

5.4.3 1996-97 Sample Selection for new institutions in cycle 2

As stated in Section 5.4.1, in cycle 1 a systematic sample of institutions was taken in each stratum with probability of selection proportional to the number of beds. Given that institutions had been chosen systematically, it was then possible to select a sample of new institutions while preserving the original design and variance estimation. To do so, all new institutions were randomly sorted and added to the bottom of the list of their corresponding stratum on the 1994 list frame, just as if they were part of the original list. Then, using the 1994 random start and sampling interval of each stratum, new institutions could be selected as if they would have been chosen in the original sample. However, when using this method, fewer institutions are selected than would have been the case if the new

institutions were stratified differently. The gain of changing stratification and therefore selecting more new institutions was deemed of little value and would have implied significant modifications in the variance calculation.

For new institutions in cycle 2, a systematic sample of residents was chosen from the list of all long-term residents.

6. Data Collection

6.1 Questionnaire Design and Data Collection Method

The NPHS: Health Institutions 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 because of illness or incapacity. The proxy respondent could be a relative, a staff member or a volunteer at the institution. Proxy respondents completed 59.1% of the interviews (of the proxy interviews, 72.8% were done by relatives of the resident). A staff member from the institution provided information on each selected resident's use of medications, their height and weight, and their contact with health professionals.

6.2 Interviewing

Collection took place between December 1996 and April 1997 with additional follow-up of residents who moved continuing until June 1997. Statistics Canada interviewers conducted the interviews.

At the beginning, all institutions were contacted by telephone by senior interviewers to arrange a meeting between an 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.

Most interviews were conducted in person. The total interview took an average of 37 minutes for non-proxy and 32 minutes for proxy respondents.

6.3 Supervision and Control

All interviewers were under the supervision of a staff of senior interviewers. The seniors were responsible for ensuring that interviewers were familiar with the concepts and procedures of the survey. They periodically monitored their 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.

6.4 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.

7. Data Processing

7.1 Data Capture

The resident questionnaire (Form 6) was captured using the Optical Coding Text Recognition (OCTR) process. In this process the questionnaires are scanned with an optical reader. If a check box had at least 3% of the area marked, the information in this field was captured. If more than one box was checked for mark-one only questions, an operator looked at the scanned image to determine which was the correct value. If the correct value was not obvious, the value of the second response was taken. When the scanner was not 90% certain of a write-in response, an operator who manually viewed the scanned image on a computer screen verified it. The unique identifier and names of medications were 100% verified by the operator.

The Institution Policy Questionnaire (Form 5) was captured manually. The information was 100% verified because of the small number of records.

The programmes written for the data capture of the questionnaires prevented out-of-range values from being entered. Editing for correct flow was performed after data capture.

7.2 Editing

After completing an interview, the interviewer reviewed the questionnaire to verify that the interview had followed the correct pattern of questions throughout the questionnaire. 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.

Head office edits included the verification of the demographic variables and response codes prior to data capture. 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).

7.3 Coding

Conditions or health problems causing activity restrictions were coded based on the International Classification of Diseases, Version 9 (ICD-9) or according to the Musculoskeletal Impairment Supplementary Coding Scheme developed for the Canadian Health and Activity Limitation Survey (HALS). Drugs and medications were coded using a revised version of the Canadian Anatomical Therapeutic Chemical (ATC) Classification System.

7.4 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 residents of 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.

7.5 Weighting

Estimation in a probability sample (such as the NPHS) is based on the principle that each person in the sample "represents," besides himself or herself, several others who are not in the sample. For example, in a simple random 2% sample of the population, each person in the sample represents 50 people in the population. In the terminology used here, it can be said that each person has a weight of 50.

The weighting phase calculates the associated weight for each person. This weight appears on the microdata file and must be used to derive meaningful estimates from the survey. For example, the number of individuals who smoke daily (see question SMI6_1 in Section 9.2.1) is estimated by selecting the records referring to those individuals in the sample having that characteristic. The weights entered on those records are then summed.

Details of the method used to calculate these weights are presented in Section 10.

7.6 Suppression of Confidential Information

‘Public Use’ microdata files (PUMF) differ in many important respects from the survey ‘master’ files held by Statistics Canada. These differences result from action taken to protect the anonymity of individual survey respondents, such as grouping variables, which could identify unique individuals. For example, individual ages are on the master file, while age groups appear on the PUMF. Users requiring access to information excluded from the microdata files may purchase custom tabulations (See Section 12.2). Estimates generated will be released to the user, subject to meeting the guidelines for analysis and release outlined in Section 9 of this document.

8. Data Quality

8.1 Response Rates

Two separate response rates can be calculated for the cross-sectional NPHS: Health Institutions file; an institutional rate and an individual response rate.

The institutional response rate identifies the percentage of in-scope institutions that agreed to allow the survey to be conducted in their facility. The residents could not be interviewed without agreement from the institution. All cycle 1 institutions still in operation and all in-scope new institutions agreed to allow the survey to be conducted in their facilities. The institutional response rate is calculated by:

$$\frac{\textit{Number of selected institutions that agreed to participate}}{\textit{Total number of in-scope selected institutions}} \times 100$$

$$= \frac{213}{213} \times 100$$

$$= 100\%$$

The individual response rate identifies the percentage of selected residents from the responding institutions with whom an interview was conducted. It is calculated by:

$$\frac{\textit{Number of residents with fully or partially completed interviews}}{\textit{Total number of selected residents within the responding institutions}} \times 100$$

$$= \frac{2118}{2383} \times 100$$

$$= 88.9\%$$

Note: Multiplying the two rates together does not give a meaningful result because different numbers of residents were selected within each 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, processing methods, etc. as those actually used in the survey. The difference between the estimates obtained from the sample and those resulting from a complete count taken under similar conditions is called the *sampling error* of the estimate.

It is unavoidable that estimates from a sample survey are subject to sampling error. Sound statistical practice calls for researchers to provide users with some indication of the magnitude of this sampling error. This section of the documentation outlines the *measures of sampling error* that Statistics Canada commonly uses. Users producing estimates from this microdata file are also urged to use these measures.

The basis for measuring the potential size of sampling errors is the standard error of the estimates derived from survey results. Because of the large variety of estimates that can be produced from a survey, the standard error of an estimate is usually expressed relative to the estimate to which it pertains. This resulting measure, known as the coefficient of variation (CV) of an estimate, is obtained by dividing the standard error of the estimate (which is equal to the square root of the variance of the estimate) by the estimate itself. It is expressed as a percentage of the estimate.

For example, suppose that, based on the survey results, one estimates that *10.4%* of residents of health institutions are daily cigarette smokers and that estimate yields a standard error of *.0094*. The coefficient of variation of the estimate is then calculated as:

$$\left(\frac{.0094}{.104}\right) x\,100\% = 9.04\%$$

For more information on the variance calculation for this survey, see Section 11. For interpretation of cv's and release guidelines see section 9.4.

8.2.2 Non-Sampling Errors

Errors not related to sampling may occur at almost every phase of a survey operation. Interviewers may misunderstand instructions, respondents may make errors in answering questions, the answers may be incorrectly entered on the questionnaire and errors may be introduced in the processing and tabulation of the data. These are all examples of *non-sampling errors*.

Over numerous 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 were spent to reduce non-sampling errors in the survey. Quality assurance measures were carried out 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 the questionnaire, and observation of interviewers to detect the misunderstanding of instructions.

Non-response to the survey is a major source of non-sampling error. The extent of non-response varies from partial non-response (failure to answer just one or some questions) to total non-response. Total and partial non-response to the institutional component of the NPHS was small. Partial non-response occurred when the respondent refused to answer a question or could not recall the requested information. Total non-response occurred because the interviewer was unable to contact the proxy-respondent or because of refusal to participate in the survey at the individual level. Total non-response was handled by adjusting the weight of the residents who responded to the survey to compensate for those who did not respond.

8.3 Imputation

Imputation was used to derive scores for one variable in the Health Institutions component of the NPHS. The variable HSI6DHSI denotes the resident's Health Status Index (HSI) score. This measure of overall health assesses vision, hearing, speech, mobility, dexterity, emotions, cognition and pain. Overall HSI scores ranging from zero to one are calculated based on responses to a series of health status questions[2]. A complete HSI score could not be calculated if one or more of the components were not answered. At least one component of the HSI was missing for 26.5% of all respondents. Because of this high

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

level of non-response to health status questions, a decision was made to impute the HSI for the 1996-97 Health Institutions Survey. A form of **hot deck imputation** was used to impute values for the missing components so that an overall HSI could be computed for these cases.

It should be noted that the strategy used was not the same as that used in 1994. However, to preserve consistency, the 1996-97 strategy was used to impute values on both the 1996 cross-sectional file and the 1994-96 longitudinal file.

The HSI was computed based on items in 8 sub-categories from the Health Status Section of the questionnaire. A sub-core was calculated for each of these sub-sections and then further computations were carried out on these sub-scores to derive the overall HSI. Imputation was carried out on these 8 sub-scores and not on actual raw items on the questionnaire. After imputation, the HSI derived variable program was modified slightly to take as input 8 imputed values: vision, hearing, speech, mobility, emotion, cognition, dexterity, pain

Imputation was carried out in three stages:

- First there was deterministic imputation. In some cases, even though there was some non-response in the items that feed in the sub-score, there was enough information to derive the sub-score with certainty. Using this partial information, a sub-score was assigned where it was deemed appropriate to do so.

- In the second stage, hot-deck donor imputation was used to impute missing sub-scores. Records with valid values for each of the 8 sub-scores were analysed using regression techniques to determine what other variables from the questionnaire were related to the sub-score values, i.e., could be used in predicting missing sub-scores values. The related variables were used as matching variables in finding donor records to impute records with missing sub-score values.

- Finally, in some cases, it is possible that a sub-score was imputed that contradicted partial information on the recipient record. For example, the partial information on a record might indicate that the sub-score for the hearing section must be between 2 and 3. If the imputed value was outside this range, the closest value was reassigned to the imputed value within the accepted range. To continue with the example, if the imputed value was over 3, it was set to 3 and if the imputed value was less than 2, it was reset to 2.

9. Guidelines for Tabulation, Analysis and Release

The following guidelines should be followed when tabulating, analysing, publishing or otherwise releasing any data derived from the survey microdata files. Whit the aid of these guidelines, users of microdata should be able to produce figures that are in close agreement with those produced by Statistics Canada and, at the same time, will be able to develop currently unpublished figures in a manner consistent with these established guidelines.

9.1 Rounding Guidelines

The following guidelines should be followed when rounding estimates derived from the microdata files:

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

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

c) Averages, proportions, rates and percentages should be computed from unrounded components (i.e. numerators and/or denominators) and then rounded themselves to one decimal using standard rounding. In standard rounding to a single digit, if the final or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is increased by 1.

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

e) In instances where, due to technical or other limitations, a rounding technique other than standard rounding is used resulting in estimates to be published or otherwise released which differ from corresponding estimates published by Statistics Canada, users are urged to note the reason for such differences in the publication or release document(s).

f) Under no circumstances are unrounded estimates to be published or otherwise released by users. Unrounded estimates imply greater precision than actually exists.

9.2 Sample Weighting Guidelines for Tabulation

The sample design used for the NPHS Institutional component was not self-weighting. The sampling weights are not identical for all individuals in the sample. When producing simple estimates, including the production of ordinary statistical tables, users must apply the sampling weight.

If proper weights are not used, the estimates derived from the microdata files cannot be considered representative of the survey population, and will not correspond to those produced by Statistics Canada.

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

9.2.1 Definitions of types of estimates: Categorical vs. Quantitative

Two main types of point estimates of population characteristics can be generated from the microdata file for the NPHS Institutional component.

Categorical Estimates:

Categorical estimates (also referred to as estimates of an aggregate) are estimates of the number or percentage of the surveyed population possessing certain characteristics or falling into a defined category. The number of individuals who smoke daily is an example of such an estimate.

Example of Categorical Question:

SMI6_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, and other measures of central tendency of quantities based upon some or all of the members of the surveyed population. They also specifically involve estimates of the form $\hat{Y} / \hat{X}$ where $\hat{Y}$ is an estimate of surveyed population quantity total and $\hat{X}$ is an estimate of the number of persons in the surveyed population contributing to that total quantity.

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

Example of Quantitative Question:

SMI6_3: How many cigarettes do you (does ...) smoke each day now?
|_|_| Number of Cigarettes

9.2.2 Tabulation of Categorical Estimates

Estimates of the number of people with a certain characteristic can be obtained from the microdata 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 ($\hat{X}$), then

c) Dividing the numerator estimate by the denominator estimate.

9.2.3 Tabulation of Quantitative Estimates

Estimates of quantities can be obtained from the microdata file by multiplying the value of the variable of interest by the weight. This quantity is then summed over all records of interest. For example, to obtain an estimate of the *total* number of cigarettes smoked each day by individuals who smoke daily, multiply the value reported in question SMI6_3 by the weight for the record, then sum this value over

all records with a response of 'daily' to SMI6_1.

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

a) Estimate the total number of cigarettes smoked per day by individuals who smoke daily as described above,

b) Estimate the number of daily smokers by summing the weights of all records with a response of 'daily' to SMI6_1, then

c) Divide estimate (a) by estimate (b).

9.3 Guidelines for Statistical Analysis

The Health Institutions component of the NPHS has a two-stage sampling design, where institutions are selected without replacement. Using data from such surveys presents problems to 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. However, the meaning or definition of the weight in these procedures is not appropriate in a sample survey framework. Typically the estimates produced by the packages are correct, but the calculated variances 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), the results produced by the standard packages will be more reasonable. They still 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 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

The number of sampled residents contributing to the calculation of the estimate should be determined before releasing and/or publishing any estimate from the microdata file. 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 then determine the coefficient of variation (CV) of the estimate by using the SAS or SPSS variance estimation program provided (see Section 11.1) and following the release guidelines below.

Sampling Variability Guidelines

Reliability/ Quality of Estimate	C.V. (in %)	Guidelines
1. Acceptable	0.0 - 16.5	Estimates can be considered for general unrestricted release. Requires no special notation.
2. Marginal	16.6 - 33.3	Estimates can be considered for general unrestricted release but should be accompanied by a warning cautioning subsequent users of the high sampling variability associated with the estimates. Such estimates should be identified by the letter M (or in some other similar fashion).
3. Unacceptable	greater than 33.3	Statistics Canada recommends not to release estimates of unacceptable quality. However, if the user chooses to do so, then estimates should be flagged with the letter U (or in some other fashion) and the following warning should accompany the estimates: "The user is advised that . . .(specify the data) . . . do not meet Statistics Canada's quality standards for this statistical program. Conclusions based on these data will be unreliable and most likely invalid. These data and any consequent findings should not be published. If the user chooses to publish these data or findings, then this disclaimer must be published with the data."

The CV is defined as the standard error (equal to the square root of the variance of the estimate), multiplied by 100 and divided by the estimate. See Section 11 for more details on calculating the variance.

10. Weighting

The weights given to responding members of the institutional survey were based on the probability of selecting the individual as well as any adjustments for the death of selected residents during the cycle 2 collection period and non-response at the individual level. The 1996 weighting procedure is based significantly on that of 1994. For cycle 1 institutions, the 1994 final institutional weights are taken as a starting point. For new institutions, this probability of selection must be calculated. In the following section, a brief description of the 1994 weighting procedures still relevant for weighting in 1996 is included. A full description of the 1994 procedures may be found in the 1994-95 NPHS: Health institutions Public Use Microdata File documentation.

10.1 Probability of selection for 1994-95 institutions

Notation:

M_h = number of beds in stratum *h* (from list of hospitals and residential care facilities);
$M_{h,i}$ = number of beds in stratum *h*, institution *i* (from list of hospitals and residential care facilities); and
n_h = number of institutions to be selected from (size) stratum *h*.

Institutions were selected from the 1994 frame with probability proportional to the number of beds. Therefore, the probability of selecting an institution *i* in most cases was:

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

When a head office was selected (see Section 5.4.3 for more details) the probability was:

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

where $P_{h,i,j}$ was the probability that an institution *j* under the authority of head office *i* is selected. For the largest institution under *i*, $P_{h,i,j}=1$. For the other *j*'s

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

where *i'* consists of all of the institutions under head office *i*, excluding the largest one.

10.1.1 1994-95 Institutional Weight Calculations and Adjustments

At this point, initial weights can be calculated at the institutional level. However, there may be non-response at that level. Adjustments have to be made to account for those institutions that do not respond.

10.1.2 1994-95 Initial Institutional Weights

Institutional weights correspond to the number of institutions represented by the sampled institution. The **initial institutional weight** is equal to the inverse of the probability of selecting the institution.

10.1.3 1994-95 Institutional Non-response Weight Adjustment

If interviewing did not take place at a selected in-scope institution, then an adjustment is made to the other institutions within the same size stratum to account for the non-responding institution. This adjustment is equivalent to:

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

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

10.2 Probability of selection of new institution in 1996-97

As described in Section 5.4.3, new institutions were selected according to the original design. Therefore, the probability of selecting a new institution *i* was:

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

where in this case $M_{h,i}$ represents the number of beds in stratum *h*, institution *i* coming from the 1996 frame. With this selection procedure, eight new institutions were selected in cycle 2. Of these, two institutions were found to be out of scope (only providing short term care); the others agreed to participate in the survey. Taking the inverse of the probability of selecting the institution gives the **final institutional weight.** Note that the new institutions were not grouped with cycle 1 institutions when adjusting for cycle1 institutional non-response.

10.3 Probability of selecting a resident

10.3.1 Probability of selecting a resident within 1994-95 institutions

Two main objectives must be achieved when selecting the new sample of residents. One is sample representativity while the other is efficiency of the sampling method. By representativity we mean that every resident must have a probability greater than zero of being selected, and this must include new residents. In order for the sample to be efficient, that is, to have low variance estimates, it is desirable that the sampling weights of selected residents within an institution be as similar as possible. Moreover, the sample should represent the distribution of the population in that institution with respect to newly admitted residents versus residents present when cycle 1 was conducted. Simply replacing cycle 1 selected residents no longer in an institution by a sample of newly admitted residents can lead to large differences in sampling weights. As well, depending on whether the attrition rate within an institution differs considerably from the one observed at the sample level in this institution, one may have to select some residents present at the time of cycle 1. For each institution, the 1994 sampling fraction was applied to both subpopulations to determine how many extra residents needed to be interviewed. In doing so, the sampling weights of selected residents were made as similar as possible within an institution.

Notation:

$R_{h,i,b}$ = actual number of long-term residents in stratum *h*, institution *i* admitted before April 1, 1995;

$r_{h,i,b}$ = number of residents allocated to be selected from stratum *h*, institution *i* among residents admitted before April 1, 1995 (excluding longitudinal panel members);

$R_{h,i,a}$ = actual number of long-term residents in stratum *h*, institution *i* admitted on or after April 1, 1995;

$r_{h,i,a}$ = number of residents allocated to be selected from stratum *h*, institution *i* among residents admitted on or after April 1, 1995; and

r_{panel} = number of longitudinal panel member still residing in institution *i* of stratum *h*.

The probabilities of selecting a resident are:

$$\frac{r_{h,i,b} + r_{panel}}{R_{h,i,b}}$$ if selected resident was admitted before April 1, 1995 or for longitudinal panel member

or

$$\frac{r_{h,i,a}}{R_{h,i,a}}$$ *if selected resident was admitted on or after April 1, 1995.*

10.3.2 Probability of selecting a residents within new institutions

Notation:

$R_{h,i}$ = actual number of long-term residents in stratum *h*, institution *i;* and
$r_{h,i}$ = number of residents allocated to be selected from stratum *h*, institution *i*.

For new institutions each resident had an equal probability of selection given by:

$$\frac{r_{h,i}}{R_{h,i}}$$

10.3.3 Initial Personal weights

An initial personal weight can be calculated as the final institutional weight multiplied by the inverse of the probability of selecting a resident within the institution.

10.3.4 Death adjustment

Because in some cases there was a long gap between picking the sample and starting surveying, the number of deaths found during surveying was much higher than anticipated. Since the pattern of deaths of residents of health institutions is not random (less-healthy people are more likely to die), bias can be introduced in the estimates if no adjustment is done to alleviate this problem. In order to compensate for the death of selected residents during the collection period, death adjustment factors were derived and applied to the initial personal weights of responding units.

Many steps led to the creation of these death adjustment factors. First, using the longitudinal panel, the characteristics of those who died within six months of the cycle 1 interview were determined. Classes were then formed that consist of groupings of units (i.e., residents) that share the same propensity to be deceased within six months of the cycle 1 interview. Classes were formed using a clustering algorithm that arranges the sample units into a tree structure by successively splitting the data set into "branches" based on the units characteristics. The software *KnowledgeSeeker IV for Windows,* developed by ANGOSS Software International Limited, was used to generate the tree structure. We used the CHAID (Chi-square Automatic Interaction Detection) algorithm available in KnowledgeSeeker to identify at each node the characteristic that best splits the sample into groups that are dissimilar with respect to the "died within six months"/ "did not die within six months" indicator. For each class, a factor defined as the ratio of the number of residents who died within six months of the cycle 1 interview to the number of residents who did not die within six months was derived. This factor is given by:

$$\frac{\textit{sum of 1994 weights for panel members within the class who died within six months of cycle 1 interview}}{\textit{sum of 1994 weights for panel members within the class who did not die within six months of cycle 1 interview}}$$

Since classes were defined using characteristics from cycle 1, corresponding characteristics available for cycle 2 respondents are used to define membership in a class. The hypothesis made here is that these ratios, which were calculated using cycle 1 data, should be very similar for cycle 2 data. Ideally, the sum of weights of selected residents who died during the collection period should be redistributed within each stratum/class. However, in doing so, this would create unstable adjustments due to the small number of records in some stratum/class combinations. To respect as much as possible the original stratification and to produce more stable adjustments, the sample was divided into the following six poststrata: four for facilities for the aged, one for cognitive facilities and one for other rehabilitative facilities. For facilities for the aged, each region has four or five size strata. The four poststrata were formed to be approximately the same size (in terms of total weights and number of records) and generally correspond to small, medium, large, and very large institutions. It should be noted that the majority of selected residents who died came from facilities for the aged.

The sample was further divided into a poststratum/class category for every cycle 2 respondent. Then, the sum of initial personal weights of respondents was calculated for each poststratum/class category. Multiplying that sum by the ratio

of "died within six months"/ "did not die within six months" (as defined above) of the corresponding class gives the initial weighted number of selected residents who died to be allocated in that category. This initial allocation is then calibrated to the corresponding sum of initial personal weights of selected residents who died in that poststratum. The final death adjustment factor in a poststratum/class category is given by:

$$\frac{\left(\begin{array}{c}\text{sum of initial personal weights of respondents within poststratum/class category} \\ + \\ \text{sum of initial personal weights of selected residents who died allocated within poststratum/class category}\end{array}\right)}{\text{sum of initial personal weights of respondents within poststratum/class category}}$$

Each death adjustment factor was applied to the respondents initial personal weight of that poststratum/class category. After this step, only responding and non-responding units were kept on the file.

10.3.5 Personal Non-response Weight Adjustment

In some cases certain selected residents did not respond to the questionnaire. An additional adjustment has to be made to the initial personal weights to compensate for the non-respondents. As described in Section 5.1.1, the cycle 2 sample is composed of longitudinal panel members and newly selected residents. Since additional information such as date of birth and sex is available for all longitudinal panel members, the personal non-response adjustment is different for this portion of the sample (which represents about 60 % of the sample). For non-respondents coming from cycle 2 newly-selected residents, only the variable sex is available (note that this variable was imputed for five records).

The personal non-response weight adjustment for panel members is given by:

$$\frac{\text{sum of weights for respondent and non - respondent residents in an institution type/age/sex category}}{\text{sum of weights for respondent residents in an institution type/age/sex category}}$$

The weight here refers to the initial personal weight after the death adjustment is made. The age/sex grouping corresponds to the one used in the present Public Use Microdata File. Note that some of the original age/sex categories were collapsed in cognitive and other rehabilitative types of institution. The collapsing was done in order to produce a more stable adjustment.

The personal non-response weight adjustment for top-up sample is given by:

$$\frac{\textit{sum of weights for respondent and non - respondent residents in an institution type/sex category}}{\textit{sum of weights for respondent residents in an institution type/sex category}}$$

The weight here again refers to the initial personal weight after the death adjustment is made.

Multiplying the weight after the death adjustment by the personal non-response weight adjustment gives the **final personal weight** that appears on the file. Since this survey focussed on individuals and not the institutions themselves, the final institutional weight does not appear on the file.

11. Variance Calculations

The institutional component of the NPHS uses a well-known, simple variance formula to compute the variances and the CVs of estimates. It assumes that institutions are selected with unequal probabilities and with replacement. In reality, the institutions were selected without replacement, that is, once selected, an institution could not be chosen a second time. This was done for operational reasons rather than for variance improvement so the impact of assuming sampling with replacement should be negligible.

A variance computation program written in SAS and SPSS is provided as part of this microdata package. This program can be used to calculate variances for means and totals. The formulas used for calculating the variances for a total Y or a ratio R=Y/X are:

$$V(\hat{Y}_{total}) = \sum_{h=1}^{N} \frac{\sum_{i=1}^{n_h} (\hat{Y}_{h,i} - \hat{Y}_h)^2}{n_h(n_h - 1)} \quad \text{and} \quad V(\hat{R}) = \frac{1}{\hat{X}^2} \sum_{h=1}^{N} \frac{\sum_{i=1}^{n_h} (\hat{Y}_{h,i} - \hat{R}_h \hat{X}_{h,i})^2}{n_h(n_h - 1)}$$

where:

$\hat{Y}_h$ is the stratum *h* estimate for a response variable *Y* based on all of the respondents in stratum *h;*

$\hat{Y}_{h,i}$ is the stratum *h* estimate for a response variable *Y* based on all of the respondents in stratum *h*, institution *i;*

N is the number of strata;

n_h is the number of sampled institutions in stratum *h*;

$\hat{X}_h$ is the stratum *h* estimate for the ratio denominator variable *X* based on all of the respondents in stratum *h*;

$\hat{X}_{h,i}$ is the stratum *h* estimate of the ratio denominator variable *X* based on all of the respondents in stratum *h*, institution *i*;

$\hat{X}$ is the overall estimate of the ratio denominator variable *X*; and

$\hat{R}_h$ is the ratio of $\hat{Y}_h / \hat{X}_h$.

11.1 Running the variance programs

Two programs (SAS and SPSS) are included with the microdata file that allows the user to calculate totals and ratios with minimal work. The programs invoke peripheral SAS and SPSS files, also included. Users should ensure file name references are consistent when using the programs and files.

11.1.1 Running the SAS variance program

The following outlines the steps of the SAS program:

STEP 1: In the data step under 'STEP 1' the user identifies the variables for which he/she wants totals and/or ratios. A missing value must be assigned for each variables (e.g. values for don't know, refusal and not stated). Please refer to the data dictionary for values to be treated as missing. For totals, a 0/1 variable is assigned to each characteristic of interest. Likewise, for ratios, a 0/1 variable is defined for both the numerator and denominator of the ratio. In the example, three totals are being computed and the 0/1 variables are defined as tot1-tot3. In the ratio example, the variables are num1-num4 and denom1-denom4 where num1/denom1 identifies a ratio requiring variance estimates. Maintain the naming convention tot1-tot*n*, num1-num*m* and denom1-denom*m* as the program automatically uses these names. In the keep statement at the end of this step, change the tot*n*, num*m* and denom*m* variables to indicate the number of totals or ratios computed.

Quantitative estimates are calculated in a similar manner. The only difference is that the 0/1 variable is replaced by a quantity variable, where the quantity represents the value of the characteristic for the respondent.

STEP 2: In the proc format statement, the user can assign descriptive names to replace the totals and ratio names generated by the program. "Totfrmt" defines the names for totals and "ratfrmt," the names for ratios. For example, in totfrmt, 2='Popn 65+' since tot2 is calculating the estimate and variance of the total population aged 65 and over.

STEP 3: In this step, the user simply has to change the array references so that the correct number of totals, numerators and denominators are shown. For example, if the user has defined two totals and three ratios in STEP 1 then the array statements would then read:

```
array totarray{*}tot1-tot2;
array numarray{*}num1-num3;
array denarray{*}denom1-denom3;
```

The program can then be run like any other SAS program.

Note: The program calls a different subroutine when calculating totals or ratios. If only totals or only ratios are calculated it is not necessary to run both subroutines. At the end of the program there are two lines:

```
%totals;
%rates;
```

To save time the unnecessary subroutine can be "commented out" by surrounding the appropriate statement with /* and */. For example, if only totals are being calculated then:

```
%totals;
 /*   %rates;  */
```

Only the subroutine associated with variances for totals will be called.

In this example, it is not necessary to define the num1-num*m* and denom1-denom*m* variables in STEP 1 since no rates are calculated. Likewise, if the %totals line is "commented out," the tot1-tot*n* variables are not required. Simply remove the reference to these variables from the keep statement in STEP 1.

11.1.2 Running the SPSS variance program

The SPSS variance program was written in a similar fashion as for the SAS program. The program includes the same three steps as that used in the SAS program. However, for technical reasons, the order of the steps do not appear sequentially in the program as it is the case in the SAS program. But each step is properly identified within the program. The following outlines the steps of the SPSS program:

STEP 1: In the data step under 'STEP 1' the user identifies the variables for which he/she wants totals and/or ratios. A missing value must be assigned for each variable (e.g., values for don't know, refusal and not stated). Please refer to the data dictionary for values to be treated as missing. For totals, an indicator variable is assigned to each characteristic of interest. Likewise, for ratios, an indicator variable is defined for both the numerator and denominator of the ratio. In the example, three totals are being computed and the indicator variables are defined as tot1-tot3. In the ratio example, the indicator variables are num1-num4 and den1-den4 where num1/den1 identifies a ratio requiring variance estimates. Maintain the naming convention tot1-tot*n*, num1-num*m* and den1-den*m* as the program automatically uses these names.

Quantitative estimates are calculated in a similar manner. The only difference is that the indicator variable is replaced by a quantity variable, where the quantity represents the value of the characteristic for the respondent.

STEP 2: The user can assign labels to replace the totals and ratios names generated by the program. The comment instructions /*<<< LABELS FOR TOTALS <<< and /*<<< LABELS FOR RATIOS <<< indicates the location within the program where labels for totals and ratios must be specified.

STEP 3: In this step, the user simply has to change the value of macro variables !ntotal, !nrates and !dim so that the correct number of totals and ratios are shown. For example if the user has defined two totals and three ratios to be calculated, then the SPSS statements would then read:

```
!let  !ntotal=2.
!let  !nrates=3.
!let  !dim=6.
```

The macro variable !dim must be equal to twice the number of ratios to be calculated.

The name of the sub-directory where the files are stored must be changed by the user in the macro variable !let !folder.

Running the programs using the variables in the example produces output similar to this:

Estimates, Variances and CVs of Totals 1

OBS	DESCRIPTION	TOTAL	VARIANCE	CV
1	Total Popn	222967.15	29161285.48	2.42
2	Popn 65+	185139.44	26768521.00	2.79
3	Women 80+	103702.22	19183726.64	4.22

Estimates, Variances and CVs of Ratios 2

OBS	DESCRIPTION	RATIO	VARIANCE	CV
1	Close Staff Members 65+	0.40613	0.00044143	5.17
2	English 65+	0.53973	0.00037237	3.58
3	French 65+	0.17786	0.00018277	7.60
4	English/French 65+	0.10223	0.00015822	12.30

12. File Usage

12.1 Variable Naming Convention

In 1996-97, the 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 mandatory: restrict variable names to a maximum of 8 characters for ease of use by analytical software products; identify the survey cycle (1994-95, 1996-97, 1998-99...) 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 1994-95 and 1996-97. The variable names about smoking should only differ in the year position in the name that identifies the particular survey cycle in which they were collected. This convention will be followed throughout the longitudinal survey, and will be adopted by all NPHS components: the household survey, the institutional survey, the Northern 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 variable DHI6GAGE:

DH:	found in the Demographic and Household content section of the questionnaire;
I:	questions that are on the Institutions survey;
6:	appeared in 1996-97 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 1994-95, access to services in 1996-97)
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 1994-95
6 1996-97
8 1998-99
0 2000-01
2 2002-03
A 2004-05
B 2006-07
C 2008-09
D 2010-11
E 2012-13
F 2014-15

12.1.5 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

Microdata files must meet stringent security and confidentiality standards required by the Statistics Act before they are released for public access in order to protect the confidentiality of the respondents participating in the survey. To ensure that these standards have been achieved, each microdata file goes through a formal review process to ensure that an individual cannot be identified. Rare values in variables that may lead to identification of an individual are suppressed on the file or are collapsed to broader categories so that individual disclosure is minimized. Frequently, these are the variables that are the most critical for doing a complete and comprehensive analysis of the survey data. Since a significant amount of resources is spent on collecting these data, ensuring that the microdata files reach their full analytical potential is important for a complete return on the statistical investment.

Custom tabulations may also be 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. The results are screened for confidentiality and reliability concerns before release. There is a charge for this service.