JOINT CANADA/UNITED STATES SURVEY OF HEALTH

PUBLIC USE MICRODATA FILE USER GUIDE

Statistics Canada
United States National Center for Health Statistics

June 2004

Table of Contents

1. INTRODUCTION 1

2. BACKGROUND 2

3. OBJECTIVES 4

4. SURVEY CONTENT 5

5. SAMPLE DESIGN 7

5.1 Target Population 7
5.2 Sample Design and Allocation 7
5.3 Household Sampling 7
5.3.1 Sampling Frame in Canada 7
5.3.2 Sample Frame in the United States 8
5.4 Selection of Respondents 9

6. DATA COLLECTION 11

6.1 Questionnaire Design and Data Collection Method 11
6.1.1 Translations 12
6.2 Interviewing 12
6.3 Minimising Non-response 13

7. DATA PROCESSING 14

7.1 Editing 14
7.2 Coding 14
7.3 Creation of Derived and Grouped Variables 14
7.4 Weighting 15
7.5 Suppression of Confidential Information 15

8. WEIGHTING 17

8.1 Adjustments Applied to the Initial Weight 17

9 DATA QUALITY 22

9.1 Response Rates 22
9.1.1 Canadian Response Rates 22
9.1.2 United States Response Rates 23
9.2 Survey Errors 24
9.2.1 Non-sampling Errors 24
9.2.2 Sampling Errors 25
9.2.3 Variance Estimation 25
9.2.3.1 Bootstrap Method for Variance Estimation 25
9.2.3.1.1 Bootvar Program for Variance Estimation 26
9.2.3.2 Taylor Series Method (with SUDAAN) for Variance Estimation 26

10. GUIDELINES FOR TABULATION, ANALYSIS AND RELEASE 28

10.1 Rounding Guidelines 28
10.2 Sample Weighting Guidelines for Tabulation 29
10.2.1 Definitions: Categorical Estimates, Quantitative Estimates 29
10.2.2 Tabulation of Categorical Estimates 30
10.2.3 Tabulation of Quantitative Estimates 30
10.3 Guideline for Statistical Analysis 31
10.4 Release Guidelines 31

11. FILE USAGE 33

11.1 Use of Weights 33
11.2 Variable Naming Convention 33
11.2.1 Variable Name Component Structure in the JCUSH 33
11.2.2 Positions 1-2: Questionnaire Section Names 34
11.2.3 Position 3: The JCUSH Identifier 34
11.2.4 Position 4: Cycle 34
11.2.5 Position 5: Variable Type 35
11.2.6 Positions 6-8: Variable Name 35
11.3 Access to Data 36

1. Introduction

The Joint Canada/United States Survey of Health (JCUSH) is a collaborative project undertaken by the Health Statistics Division of Statistics Canada and the National Center for Health Statistics (NCHS) of the United States Centers for Disease Control and Prevention. The project called for a one-time telephone survey to be conducted in both countries, controlling for survey design and implementation effects, in order to produce comparable results.

The jointly developed survey instrument was designed with questions covering chronic conditions, functional status, determinants of health, and health care utilisation. The questionnaire was administered to approximately 3,500 Canadians and 5,000 Americans living in households. Plans called for a comparable Canada/United States random digit dial (RDD) sample design and for Statistics Canada interviewers to collect the data for both the United States and Canadian respondents.

This Microdata File contains data collected from both samples beginning November 4, 2002 and ending March 31, 2003 (collection was extended several weeks for only U.S. data in April and June of 2003 – see section 6.2 for more details) as representative of the ten Canadian provinces, all 50 United States, and the District of Columbia. Data were collected for one adult aged 18 years or older per household from persons living in private occupied dwellings. Excluded from the sampling frame were individuals living in health care institutions, nursing homes, full-time members of the Canadian or United States Armed Forces, and residents of Northern Canada (the three Territories).

This document has been produced to facilitate the analysis of the JCUSH Public Use Microdata File, which is described in detail in the following text and related documents.

Any questions about the data or their use should be directed as follows:

For technical and general data support in the United States:
NCHS Information line: 1 (301) 458-INFO (4636) or
Toll free: 1 (866) 441-NCHS (6247)
For technical and general data support in Canada:
Data Access Unit, Health Statistics Division: 1-613-951-1653
E-mail: jcush-ecces@statcan.ca

2. Background

The Joint Canada/United States Survey of Health (JCUSH) is an outgrowth from a session at the 2000 United States/Canada Interchange on making data more comparable and integrated. The Interchanges, consisting of two-day meetings with the site of the meeting alternating between the Washington DC area and Ottawa, began in 1999. The purpose of the meetings is to promote communication, collaboration, cooperation, synergy, facilitation of comparative analyses, and interaction between countries. Participants at that 2000 Interchange suggested that the best way to make comparisons would be to conduct a joint survey in which the questionnaire, sample design, data collection methods, data processing, and editing were done at the same time in the same way. A team of staff members from the two agencies was designated to explore the feasibility of the project and implement it.

Through integrating survey design and data collection methodologies, the JCUSH attempts to increase knowledge about the comparability of the two health data systems, as well as provide a model for future international comparisons. Although previous studies have examined cross-cultural comparisons of social, economic, and political characteristics using separate data sets, a number of methodological limitations exist in their research. In a study comparing immigration trends of Canada and the United States, Pryor and Long [1] posited that comparability is hindered by the range of data available, the universes included, timing of data sources, and differing concepts and definitions. In a study of criminal convictions and sanctions in Europe, Barclay [2] found that wide differences in the levels and degrees of sanctions were due in part to different recording practices. When analyzing data from three cross-cultural surveys that differed in geographical size and range, Dow [3]found that sample design effects across data sets lead to larger variances and increased risk of Type I errors. Scheuch [4] identified the following problems of cross-cultural research and comparability using different data sets: question meaning, equivalence of indicators, unit of analysis, and administrative problems unique to each study.

[1] Pryor, Edward T. and John F. Long (1987). "The Canada-United States Joint Immigration Study: Issues in Data Comparability." *Workshop on International Migration Data,* Ottawa, Canada.

[2] Barclay, Gordon C. (2000). "The Comparability of Data On Convictions and Sanctions: Are International Comparisons Possible?" *European Journal On Criminal Policy and Research* 8*:* 13-26.

[3] Dow, Malcolm M. (1993). "Saving the Theory: On Chi-Square Tests With Cross-Cultural Survey Data." *Cross-Cultural Research* 8*:*3,4*:* 247-276.

[4] Scheuch, Erwin K. (1993). "The Cross-Cultural Use of Sample Surveys: Problems of Comparability." *Historical Social Research* 18: 104-138.

Differences in survey results may exist even when taking into account methodology specific to each body of data. In a study of child care data in United States surveys, Raley et al.[5] found that when controlling for effects particular to individual studies, the methodological differences (time frame of the study, differing methods of screening respondents) continued to produce substantial differences in results. While debates such as this over validity of comparative research continue, the JCUSH adds to the literature as well as produces a body of fully comparable data between Canada and the United States.

[5] Raley, et al. (2000). "The Quality and Comparability of Child Care Data in United Sates Surveys." *Social Science Research* 29: 356-381.

3. Objectives

The principal objectives of the JCUSH are:

- To develop, implement, and document a collaboration between national statistical offices for conducting joint health surveys of their national populations;
- To use knowledge gained in conducting the JCUSH to modify or fine-tune questionnaires from the two countries' ongoing national health surveys so as to enhance comparability between those surveys; and
- To produce a data set with highly comparable data on the Canadian and United States populations for use by researchers studying the effect of variations in health systems, health care, health status, and functional status, and for use in survey metholodogical studies.

4. Survey Content

This section provides a general summary of the content selected for inclusion in this survey. The content for the JCUSH is based on portions of the existing questionnaires from the two nations' separate ongoing health surveys: the National Health Interview Survey (NHIS) in the United States and the Canadian Community Health Survey (CCHS) and National Population Health Survey (NPHS) in Canada. Topics of study included in the JCUSH are summarized in Table 4.1.

Table 4.1 Questionnaire Modules

General Health	Limitation of Activities (Random Block from the United States)
Restriction of Activities (Random Block from Canada)	PAP Smear Test
Chronic Conditions	Mammography
Depression	Dental Visits
Contact with Mental Health Professionals	Insurance
Smoking	Vocational Restriction of Activities (Random Block from the United States)
Health Utility Index (Random Block from Canada)	Patient Satisfaction
Height and Weight	Physical Activities
Health Care Utilisation	Socio-demographic Characteristics
Use of Medications	

The United States and Canadian ongoing health surveys are relatively similar in regards to content; that is, they cover the same range of health issues — chronic and acute conditions, mental and physical health, medical services and health care coverage, etc. They differ primarily with respect to their measurement of particular health entities; that is, they may conceptualize an entity differently and therefore ask different questions about it. This is particularly true for what might be called, in a general sense, "functional status."

Each survey employs different approaches to measuring functional status. The Canadian surveys employ a set of questions covering "Restriction of Activities," and the "Health Utility Index," which was developed at Canada's McMaster University. The NHIS employs both a measure of limitation in major role activities and a set of functional limitations questions. All of these measures are clearly related to one another, both on the face of it and in empirical analyses, but they produce different estimates of the levels and distributions of functional status in populations. All four of these measures are included in the JCUSH to address these differences, and to provide researchers with the opportunity to examine the performance of each of these measures on two populations. The four functional status modules are denoted with the words, "Random Block" (Table 4.1) to indicate that all four modules were administered to each respondent, but in a random sequence. This was done in an attempt to reduce possible response bias that may have arisen from respondent fatigue or impatience with answering what may appear to be some of the same questions repetitively.

The functional status questions and most of the other questions were asked of both Canadian and United States respondents. However, due to differing procedural and confidentiality requirements of each country, as well as differences between the countries themselves, certain groups of questions were administered only to Canadian or United States respondents, and others cannot be compared directly without adjustment factors. These include the following:

A. Separate sets of racial or cultural background questions were asked of Canadian and United States respondents.

B. Due to differing health care systems, various health insurance coverage questions were asked of United States respondents only.

C. Cross-country income comparisons cannot be made without adjusting for currency differences between the two countries. To make income comparisons possible, two exchange rates are provided in the data. The first converts Canadian to United States dollars, and the second converts United States to Canadian dollars. As exchange rates fluctuate daily, the exchange rates provided in the data are the medians of the daily exchange rates that occurred during data collection. To circumvent currency conversion problems, income quintiles are also included in the data. These variables indicate where in the relative income distribution of each country a respondent's or household's reported income falls and are directly comparable across countries without any need for adjustments. Please refer to the Derived and Grouped Variables Documentation for more details on theses variables.

5. Sample Design

5.1 Target Population

The target population of the JCUSH is Canadian and United States household residents aged 18 years or older. The institutionalized population is excluded, as are people in prison and full time members of the Canadian or United States Armed Forces. In Canada, the three northern territories (Yukon, Northwest Territories and Nunavut) were excluded. Similarly, in the United States, the United States territories (Puerto Rico, the United Sates Virgin Islands, American Samoa, Guam and the Commonwealth of the Northern Mariana Islands) were excluded, but residents of the District of Columbia were included.

5.2 Sample Design and Allocation

The JCUSH sample was designed to produce reliable national estimates for three age groups (18 to 44, 45 to 64 and 65 years and older), by sex. Statistics Canada and NCHS were each responsible for designing their respective samples. To provide reliable national estimates for three age groups, by sex, and to adhere to the budget allocated to the JCUSH, a sample of 3,500 respondents in Canada and 5,000 respondents in United States was desired. These sample sizes were enlarged before data collection to take into account out-of-scopes and anticipated non-response.

The JCUSH sample was stratified by province in Canada and by four geographic regions in the United States (Northeast, Midwest, West and South). In each country, the sample was proportionally allocated within each stratum based on their population sizes.

5.3 Household Sampling

Several approaches to select the sample were considered but the need for having one method applicable to both countries limited the options. The sample selection method allowing for the best comparability between the two countries was Random Digit Dialling (RDD). Each organization was responsible for drawing its own sample.

5.3.1 Sampling Frame in Canada

The sampling of households from the RDD frame uses the Elimination of Non-Working Banks (ENWB) method.[6] The frame begins as a list of all possible "hundreds banks," each of which consists of 100 consecutive telephone numbers that share the same first eight digits of the ten-digit telephone number. For example, the numbers from 617-555-7100 to 617-555-7199 constitute a bank of 100 telephone numbers that share the same first eight digits. By using all active area codes in Canada and all active prefixes within those area codes, the frame begins with

[6] Norris, D.A. and Paton, D.G. (1991*). Canada's General Social Survey: Five Years of Experience*. Survey Methodology, 17, 227-240.

all possible "hundreds banks." Next, a bank is classified as non-working if it does not contain any
residential telephone numbers, and, as non-working banks are identified, they are eliminated from the frame. It should be noted that these banks are eliminated only when there is evidence from various sources that they are non-working. When there is no information about a bank, it is left on the frame. The banks remaining on the frame are grouped to create RDD strata. Within each RDD stratum, a bank is randomly chosen (e.g., 617-555-71XX), and the final two digits of the telephone number are generated at random (e.g., a number between 00 and 99 is selected) to create a complete, ten-digit telephone number. This procedure is repeated until the required number of telephone numbers within the RDD stratum is reached. Frequently, the number generated is not in service or is out-of-scope, and therefore many additional numbers must be generated to reach the targeted sample size. This success rate is referred to as the *hit rate* and varies from region to region. Within the JCUSH, Canadian hit rates ranged from 43% to 58% at the provincial level.

5.3.2 Sample Frame in the United States

In the United States, the JCUSH employed a list-assisted RDD sample frame.[7] The list-assisted method uses prefix area combinations of area codes and central office codes as the basis of constructing a sampling frame of banks of 100 consecutive telephone numbers (e.g., 301-937-5900 to 301-937-5999). The RDD sample was selected using the GENESYS Sampling System (a proprietary product of Donnelley Marketing Information Systems (DMIS)). Following creation of these 100-number banks, GENESYS identifies banks that have a very low probability of containing working residential numbers. These numbers are deleted from the sampling frame. From the retained banks of 100 numbers, known as the 1+ directory-listed residential telephone numbers, a random sample of complete ten-digit telephone numbers is drawn in such a way that each number has a known and equal probability of being selected.

The GENESYS system incorporates a capability for screening out a portion of the non-working numbers as a preliminary sample preparation activity. The screening is implemented by DMIS in three phases. In the first phase, the sample is matched against a list of directory-listed residential telephone numbers: any such numbers are retained for the third phase. The remaining sample is also matched against a list of business numbers from the Yellow Pages to exclude business numbers. In the second phase, a system called GENESYS-ID screens the remaining sample to remove a portion of the non-working numbers. Using personal computers with special hardware and software, this system automatically dials all the telephone numbers to detect intercept signals that indicate non-working numbers. The third and final phase merges the file of remaining telephone numbers (after removing the business and non-working numbers) with the file of directory-listed residential numbers that were retained in the first phase. The numbers resulting from this phase were sent to the Computer-Assisted Telephone Interviewing (CATI) system.

[7] Lepkowski, J.M. (1988). *Telephone Sampling Methods in the United States*. In Telephone Survey Methodology (Eds. R. Groves et al.), 73-98. New York: John Wiley and Sons.

5.4 Selection of Respondents

As mentioned earlier, an objective of the JCUSH was to obtain reliable estimates at the national level for six domains: three age groups (18 to 44, 45 to 64 and 65 years and older) by sex. With the RDD method, it is difficult to control the sample composition since the age and the sex of the respondents are unknown beforehand. Since males aged 65 years and older represent only about 7% of the population, and since only about 13% of the households contain at least one male aged 65 years or older, a purely random selection of the respondents among the adult household members would have necessitated a very large sample size to guarantee reliable estimates for this group. For the JCUSH, the age group 65 years and older is important. To avoid an overly large sample and to respect operational and budget constraints, it was decided to increase the probability of selection for persons aged 65 years and older.

To increase the selection probability in this group, the computer application was designed to randomly select the respondent from among only the household members aged 65 years and older when at least one person in the household was part of this group. For households containing only people younger than 65 years old, the respondent was randomly selected from among all the adult members. This strategy slightly increased the representation of those 65 years and older in the sample, without creating an overly large distortion compared to the observable distribution in the population. The main inconvenience of this approach is that it systematically excludes from the sample the population younger than 65 years old living with one or more people aged 65 years and older. A bias might be introduced in the sample if these people have particular characteristics. On the other hand, this approach avoids obtaining extreme weights. Such weights would be obtained for the population younger than 65 years old living with one or more people 65 years old and older, if their probability of selection was decreased and close to zero. For this reason and to ensure a sufficient representation of those 65 years and older, it was concluded that the possible bias was an acceptable compromise.

Table 5.1 presents a comparison between the unweighted respondent distributions and the target population distributions for both countries. It also shows that males, especially those aged 18 to 44 years, are underrepresented among respondents. The fact that males, particularly in that age group, are more difficult to reach and to interview is known and is not unique to this survey.

Table 5.1 Canadian and United States unweighted respondent distributions vs. target population distributions (2002)

Population		United States Respondents (%)	United States Target Population (%)	Canadian Respondents (%)	Canadian Population (%)
18-44 years	**Males**	19.69	25.78	21.65	26.31
	Females	25.42	26.52	26.50	25.95
45-64 years	**Males**	14.92	15.35	14.64	15.90
	Females	17.68	16.34	15.92	16.21
65 years and older	**Males**	8.26	6.85	9.64	6.85
	Females	14.03	9.16	11.81	8.78

Source: 1996 Census of Population (1996) Statistics Canada, and Current Population Survey (2002) U.S. Department of Commerce, Economics and Statistics Division, Bureau of the Census.

6. Data Collection

6.1 Questionnaire Design and Data Collection Method

Both countries gained legal clearance for the JCUSH under separate protocols and operated under separate legal authorities when interviewing respondents from the two countries.

The collection of data in the JCUSH from United States respondents was authorized by Section 306 of the *Public Health Service Act*. This Act also provided the legal requirement (section 308(d)) for protecting the confidentiality of the data. Data are further protected from release that may compromise the privacy of respondents by the *Privacy Act of 1974* and updates. The JCUSH design and questionnaires were reviewed by an NCHS Institutional Review Board to ensure that the rights of respondents are maintained. Lastly, the Office of Management and Budget reviews and approves surveys to assure the public that the government is not placing an undue burden on them mainly by ascertaining that the research effort is for the public good and does not duplicate other ongoing efforts.

As part of the United States Institutional Review Process, Statistics Canada employees working on the study were required to sign pledges of confidentiality. As agents of NCHS, members of Statistics Canada staff were legally responsible for adhering to confidentiality policies and procedures of NCHS when collecting data on United States respondents.

When collecting data for Canadian respondents, Statistics Canada operated under the authority of its governing legislation, the *Statistics Act*, which gives the Agency a very comprehensive mandate, requiring it to collect, compile, analyze, abstract, and publish information on the economic, social and general conditions of the country and its citizens, and to produce statistics on a very detailed list of specific matters, including health. The Canadian component of the JCUSH was duly authorized and prescribed under section 7 of the *Statistics Act*. Further, by law, information collected under the *Statistics Act* relating to identifiable persons, businesses or organizations cannot be made available to anyone outside of Statistics Canada without their consent, nor can it be accessed under any other legislation such as the *Access to Information Act*. The obligation to protect the confidentiality of information obtained under the *Statistics Act* rests with every Statistics Canada employee, each of whom must take an Oath of Secrecy and is subject to fine and/or imprisonment for any breach of confidentiality.

The JCUSH questionnaire was administered using the Computer-Assisted Telephone Interviewing (CATI) method. CATI offers a number of data quality advantages over other collection methods. First, question text, including reference periods and pronouns, is customised automatically based on factors such as the age and sex of the respondent, the date of the interview, and answers to previous questions. Second, edits to check for inconsistent answers or out-of-range responses are applied automatically, and on-screen prompts are shown when an invalid entry is recorded. Immediate feedback is given to the respondent, and the interviewer is able to correct any inconsistencies. Third, questions that are not applicable to the respondent are skipped automatically.

6.1.1 Translations

In Canada, the Communications Policy of the Government of Canada and the *Canadian Charter of Rights and Freedoms* establish the equal status of English and French as the two official languages of Canada. The Charter enshrines the right of the Canadian public to communicate with the Government of Canada in either language. Communications with the public and services to the public must be provided in both languages as required by the *Official Languages Act*. To follow these guidelines, all survey questionnaires and related documentation are developed, produced and available at the same time in both official languages, English and French, and the two versions are equivalent.

In the United States, According to the United Sates Bureau of the Census data,[8] more than 15% of all United States households speak a language other than English. In order to achieve high response rates it is crucial that non-English-speaking households be included in the sample to the greatest extent possible. To address this issue, the JCUSH modules were translated into Spanish for United States residents, and appropriate procedures were developed (appropriate respondent help screens, question-specific probes, Spanish-speaking interviewing staff) to handle Spanish-speaking households. All staff used for interviewing Spanish-speaking households were bilingual.

For the Spanish version of the questionnaire, NCHS contracted an external firm to translate the survey. A Spanish review conference was held at NCHS in early June 2002. Although the translators were not certified, they were experienced, with extensive survey translation experience and had worked with NCHS on several occasions.

6.2 Interviewing

Data collection took place between November 4, 2002 and March 31, 2003. Additional collection took place during several weeks in April and June 2003 for only the United States sample to focus on encouraging selected persons who had previously refused to participate in the survey. In all selected households, a knowledgeable household member aged 18 years or older was asked to supply basic demographic information on all residents of the household. A household member aged 18 years or older was then randomly selected for a more in-depth interview.

In cases where the randomly selected respondent was incapable of completing an interview, another knowledgeable member of the household supplied information about the selected respondent. This is known as a proxy interview. While proxy respondents were able to provide accurate answers to most of the survey questions, the more sensitive or personal questions were sometimes beyond the scope of knowledge of a proxy respondent. This may result in some questions from the proxy interview being unanswered. Therefore, every effort was taken to keep proxy interviews to a minimum.

[8] United States Department of Commerce, Economics and Statistics Division, Bureau of the Census, 2000.

Both the Canadian and United States interviews were conducted by Statistics Canada permanent employees from Statistics Canada's Regional Offices using the same questionnaire. Interviewers are employees hired and trained specifically to carry out surveys using computer-assisted interviewing, and most are experienced interviewers. All interviewers attended a training session and received a manual for use as a reference tool. The questionnaire was administered in three languages: French and English for Canadian interviews and Spanish and English for American interviews. Interview duration was about 30 minutes.

6.3 Minimising Non-response

Prior to the first contact by an interviewer, an introductory letter was mailed to each selected dwelling for which a valid mailing address was available. This explained the importance of the survey and assured confidentiality of the respondents.

Advance letters for both countries were nearly identical, the divergence stemming from the mention of authorizing legislation (The *Canadian Statistics Act* versus the *United States Public Health Service Act*) and agencies involved. The letters were written to meet the requirements of both agencies' institutional criteria, reflecting the effort of staff from both countries to make the letter concise and readable at an 8^{th} grade level. Statistics Canada was responsible for mailing out advance letters to Canadians in the sample, while NCHS mailed the advance letters to the United States sample through the United States Public Health Service mailing facility in Rockville, Maryland.

Interviewers were instructed to make all reasonable attempts to obtain interviews. When the timing of the interviewer's phone call was inconvenient, an appointment was made to call back at a more convenient time. If no one was home, numerous call-backs were made. For individuals who at first refused to participate in the survey, a letter was sent to the respondent stressing the importance of the survey and the household's collaboration. This was followed by another call from a senior interviewer, a project supervisor or another interviewer to try to convince respondents of the importance of participating in the survey. During the final months of data collection, collection efforts focused on non-response cases and on selected persons who had previously refused to participate in the survey.

7. Data Processing

7.1 Editing

Much editing of the data was performed at the time of the interview by the computer-assisted telephone interviewing (CATI) application. It was not possible for interviewers to enter out-of-range values, and flow errors were controlled through programmed skip patterns. For example, CATI ensured that questions that did not apply to the respondent were not asked. These fields are automatically set to "not applicable," meaning that the respondent does not get asked the question because he/she is not in the category of interest (because of age, sex, marital status, etc.). For example, young people do not get asked questions on labour, and women who do not get asked questions on prostate cancer.

The "not applicable" values should not be confused with the "not stated" values. The latter include situations where the respondent/interviewee does not remember the answer and so cannot give a response, but stays on the telephone and continues the interview with the rest of the questionnaire. For example, a specific date is asked for, and the respondent/interviewee remembers the year, but not the month and the day. The month and the day will be coded as not stated. Also, if the person hangs up the telephone after answering a few questions, the rest of the questionnaire where there are no answers will be coded as not stated. An interviewee may also refuse to answer a question or may not know the specific answer to a question, in which case a value of "refused" or "don't know" is assigned.

In response to some types of inconsistent or unusual reporting, warning messages were invoked but no corrective action was taken at the time of the interview. Where appropriate, edits were instead developed to be performed after data collection at Statistics Canada's Head Office. Inconsistencies were usually corrected by setting one or both variables in question to "not stated."

7.2 Coding

Pre-coded answer categories were supplied for all suitable variables.

Several questions in the JCUSH questionnaire allow write-in responses. For some of these questions, write-in responses were coded into one of the existing listed categories if the write-in information duplicated a listed category.

7.3 Creation of Derived and Grouped Variables

To facilitate data analysis, a number of variables on the file have been derived using items found on the JCUSH questionnaire. They are called "derived variables" in Canada and "recoded variables" in the United States. Derived variables generally have a "D" or "G" in the fifth character of the variable name. In some cases, the derived variables are straightforward, involving collapsing of response categories. In other cases, several variables have been combined

to create a new variable. The Derived and Grouped Variables Documentation provides details on how these more complex variables were derived.

7.4 Weighting

The principle behind estimation in a probability sample such as the JCUSH is that each person in the sample represents himself/herself and a number of others not in the sample who have similar socio-demographic characteristics. For example, in a simple random sample in which each person had a 1/50 (or 2%) probability of being selected, each person in the sample represents 50 persons in the population. In the terminology used here, it can be said that each person has a weight of 50.

The weighting phase is a step that calculates, for each person, his or her associated sampling weight. 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 is calculated by selecting the records for individuals in the sample having that characteristic and summing the weights entered on those records. Details of the method used to calculate sampling weights are presented in Section 8.

7.5 Suppression of Confidential Information

It should be noted that the Public Use Microdata File described herein differs in a number of important respects from the survey "Master File" held by the agencies. These differences are the result of actions taken to protect the anonymity of individual survey respondents. Protection of respondents' confidentiality is assured through suppression of individual values, variable grouping, and variable capping in the Public Use Microdata File. Please refer to Table 7.1 for a complete list of variables that have been collapsed or capped on the Public Use Microdata File. The Data Dictionary and the Derived and Grouped Variables Documentation will also provide all the definitions of these derived variables.

Table 7.1 Variables Collapsed or Capped

Variable Name	Definition	Grouping or Capping
DHJ1GNHH	Total number of people in household	Capped at 5 or more
DHJ1GAGE	Age	Capped at 85 years old or more
HUJ1GDEX	Dexterity trouble - function code	Collapsed into 3 categories
HUJ1GHER	Hearing problems - function code	Collapsed into 3 categories
HUJ1GMOB	Mobility trouble - function code	Collapsed into 4 categories
HUJ1GSPE	Speech trouble - function code	Collapsed into 2 categories
HUJ1GVIS	Vision trouble - function code	Collapsed into 5 categories
HCJ1G2A	Family doctor visits	Capped at 31 or more
HCJ1G2B	Eye doctor visits	Capped at 12 or more
HCJ1G2C	Chiropractor visits	Capped at 31 or more
HCJ1G2D	Nurse visits	Capped at 12 or more
HCJ1G2E	Dentist visits	Capped at 12 or more
HCJ1G2F	Physiotherapist visits	Capped at 31 or more
HCJ1G2G	Psychologist visits	Capped at 12 or more
HCJ1G2H	Speech therapist visits	Capped at 12 or more
HCJ1G2I	Other visits	Capped at 12 or more
HCJ1GMC	Number of consultations with doctor	Capped at 12 or more
CMJ1G01L	Mental health professional contacts	Capped at 12 or more
HCJ1G01A	Overnight stays at hospitals nursing homes	Capped at 31 or more
SDJ1GMS	Marital status	Collapsed into 4 categories
SDJ1GHED	Highest level of education - respondent	Collapsed into 4 categories
SDJ1GCBC	Country of birth - Canada only	Collapsed into 2 categories
SDJ1GCBU	Country of birth - US only	Collapsed into 2 categories
IWJ1GMSI	Total household income - main source	Collapsed into 5 categories
IWJ1GTHI	Total household income - best estimate	Capped at $30,000 or more
IWJ1GTPI	Total personal income - best estimate	Capped at $80,000 or more
IWJ1GHEQ	Home Equity	Capped at -$500,000 or less and $500,000 or more

8. Weighting

In order for estimates produced from survey data to be representative of the target population, and not just of the sample itself, users must incorporate the survey weights into their calculations. A survey weight is given to each person included in the final sample, that is, the sample of persons who responded to the survey questions. This weight corresponds to the number of persons represented by the respondent for the target population.

The weights for the Canadian and the United States samples were obtained separately, but both used the method described below. Table 8.1 presents an overview of the different adjustments, part of the weighting strategy, in the order in which they were applied.

Table 8.1 List of Adjustments in the Weighting

0	RDD initial weight
1	Removal of out-of-scope numbers
2	Household non-response adjustment
3	Multiple telephone line adjustment
4	Creation of person-level weight
5	Person non-response adjustment
6	Post-stratification

8.1 Adjustments Applied to the Initial Weight

Adjustment 0 – RDD Initial Weight

Random Digit Dialing was used for this survey, which gives rise to a stratified simple random sample (without replacement) of residential telephone lines. Thus, an "RDD initial weight" is given by the inverse probability of selecting a residential telephone line from a list of telephone numbers. The RDD initial weight (different for each stratum) is given by :

$$W_{initial} = \frac{\textit{total number of telephone numbers in the sampling frame (N)}}{\textit{total number of telephone numbers that were randomly sampled from that sampling frame}}.$$

Note that some original strata in Canada were collapsed for confidentiality reasons.

Adjustment 1 – Removal of Out-of-scope Numbers

Telephone numbers leading to businesses, institutions, as well as numbers not in service, are all examples of out-of-scope cases for a telephone frame.

The weight for sampled telephone numbers found to be out-of-scope is set to 0 (using a dummy adjustment factor). In the United Sates, a certain number of telephone numbers remained unresolved at the end of the collection (see section 9.1). All the unresolved numbers were kept

after this adjustment, but their initial weight was multiplied by an adjustment factor ($P_{in\text{-}scope}$) given by the proportion of the "in-scopes" among all the resolved numbers (in-scopes or out-of-scopes). The resulting weight was obtained by multiplying the initial weight by the following adjustment, at the WTS_STR level:

$$A_1 = \begin{cases} 0 \text{ if out of scope} \\ P_{in\text{-}scope} \text{ if unresolved (United States only)} \\ 1 \text{ otherwise} \end{cases}$$

Out-of-scope records (weight=0) were thereafter dropped from the file.

Adjustment 2 – Household Non-response Adjustment

Despite all the attempts made by the interviewers, some non-response at the household level is inevitable. Non-response encompasses any of the following situations: refusal, special circumstance, language barrier, no one at home, temporarily absent, or computer problem. Non-response is compensated for by proportionally adjusting the weights of responding households. This adjustment was done at the stratum level and is given by:

$$A_2 = \frac{\text{sum of weights for all sampled households}}{\text{sum of weights for respondent households}} .$$

Again, original strata in Canada were collapsed for confidentiality reasons.

Adjustment 3 – Multiple Telephone Line Adjustment

A household that contains multiple residential telephone lines that are listed on the RDD frame has a larger probability of being selected than a household with one such line. Thus, sampled households within which it is ascertained that multiple telephone lines are present are assigned a weight adjustment equal to the inverse of the number of residential telephone lines within the household. Note that this information was obtained during the early stage of the interview. To reduce the variability of the weights, it was decided to use a maximum of 3 telephone lines for this adjustment. Each household with more than 3 telephone lines was considered to have 3 lines only. This multiplicative adjustment, done at the stratum level, is given by:

$$A_3 = \frac{1}{\text{number of residential voice telephone lines within the household (max 3)}} .$$

Note: Due to a technical problem with the computer application, the number of residential telephone lines within the household was not available for the records completed during the first two weeks of collection. When the information was not available, the weights were divided by the average number of residential telephone lines within a household, obtained using the records for which the information was available. This correction was done separately for each household size (1, 2, 3, 4 and 5 or more), since the number of people in the household has an impact on the number of telephone lines.

Adjustment 4 – Creation of Person-Level Weight

This adjustment converts the household-level weight to a person-level weight. As mentioned above, only one person aged 18 and over is selected from each sampled household. To reduce the variability of the weights, it was decided to use a maximum of 3 persons aged 18 years and over for this adjustment for the United States sample. Each United States household with more than 3 persons aged 18 years and over was considered to have 3 persons only. A multiplicative weight adjustment must be made to reflect the selection and is given by:

$$A_4 = \frac{1}{\textit{probability of selection within the household}}.$$

Note: See Section 5 for more details about the unequal probabilities of selection for persons aged 65 years and older.

Adjustment 5 – Person Non-response Adjustment

This adjustment consists of compensating for the effects of non-response at the person level. It may happen that, although a household is considered to be "responding," the information for the selected member of the household was not completed. The members for which this is true are considered to be selected member non-respondents, and a weight adjustment is made to responding selected members in the same age-sex-stratum-household size category to compensate. The household size (1, 2, or 3 or more) was used because of its correlation with the response rate at the selected member level. The multiplicative adjustment is given by:

$$A_5 = \frac{\textit{sum of weights of all sampled selected members in an age-sex-stratum-hhld size category}}{\textit{sum of weights of respondent selected members in an age-sex-stratum-hhld size category}}.$$

The age categories used for each sex were: 18 to 44, 45 to 64, and 65 and older. Some original strata in Canada were collapsed due to confidentiality restrictions.

Adjustment 6 – Post-stratification

Finally, post-stratification was done to ensure that the final weights sum to the population estimates, for some auxiliary variables. In Canada, population estimates are based on the 1996 Census of Population,[9] and in the United States, estimates were based on the October 2002 Current Population Survey.[10] The auxiliary variables used to create the post-strata are listed in Table 8.2.

Table 8.2 Auxiliary Variables

Canada	United States
Age (5 groups): 18-34 35-44 45-54 55-64 65+	Age (5 groups *): 18-34 35-44 45-54 55-64 65+
Sex: Male Female	Sex: Male Female
Region: 1 - Atlantic 2 - Quebec 3 - Ontario 4 - Prairies 5 - British Columbia	Race/Ethnicity: 1 - Hispanic 2 - Non-Hispanic / Black 3 - Non-Hispanic and Non-Black

* For Race/Ethnicity = Hispanic or Non-Hispanic/Black, only three age groups were used (18-44, 45-64, 65+) due to sample size constraints.

For both Canadian and United States portions of the sample, five age groups (with one exception in the United States) were used instead of three in order to get a better sample distribution compared to the population distribution. This also reduced the impact of the unequal probabilities of selection of the respondents (no one under 65 years old living with someone over 65 years old was selected, so this group was underrepresented in the sample).

This adjustment also takes into account the fact that a small percentage of the households do not have a telephone. Because the population estimates give the counts of all persons in the target population, regardless of whether the household has telephone service, this adjustment corrects in part for adults who belong to households that do not have telephones. It does not adjust for any biases that are introduced if non-response due to not having a telephone is not at random.

[9] 1996 Census of Population (1996). Statistics Canada. Ottawa.

[10] Current Population Survey (2002). U.S. Department of Commerce, Economics and Statistics Division, Bureau of the Census.

The multiplicative weight adjustment is given by:

$$A_6 = \frac{\textit{population estimate for a post-stratum}}{\textit{sum of the weights of respondent selected members in a post-stratum}} .$$

Final Weight

Consequently, the final weight is formed by multiplying the RDD initial weight by adjustments 1 to 6:

$$W_{final} = W_{initial} \times A_1 \times A_2 \times A_3 \times A_4 \times A_5 \times A_6 .$$

The final weight can be found on the data file with the variable name WT_SAM.

9 Data Quality

9.1 Response Rates

The overall response rates are 65.5% for the Canadian sample and 50.2% for the United States sample. A major issue caused by the presence of invalid numbers in the RDD sample is the difficulty in determining if the numbers with no answer are valid or not. In Canada, the small number of telephone companies allows the use of the companies' lists to validate the numbers. In the United States, the larger number of phone companies makes this practice impossible, which implies that the validity of several numbers remains unknown. For that reason, the response rates for Canada and the United States are calculated in accordance with different guidelines.

9.1.1 Canadian Response Rates

In total and after removing the out-of-scope units, 5,355 Canadian households were selected to participate in the JCUSH. Out of these selected households, a response was obtained for 3,858, which results in an overall household-level response rate of 72.0%. Among these responding households, 3,858 individuals (one per household) were selected to participate in the JCUSH, out of which a response was obtained for 3,505, which results in an overall person-level response rate of 90.9%. At the Canada level, this would yield an overall response rate of **65.5%**.

Household-level response rate (HHRR)

$$\text{HHRR} = \frac{\text{\# of responding households}}{\text{all in-scope households}} .$$

Person-level response rate (PRR)

$$\text{PRR} = \frac{\text{\# of responding persons}}{\text{all selected persons}} .$$

Overall response rate = HHRR x PRR .

Next is an example of how to calculate the combined response rate for Canada.

HHRR = 3,858 / 5,355 = 0.720 .
PRR = 3,505 / 3,858 = 0.909 .

Overall response rate = 0.720 x 0.909
= 0.655
= **65.5% .**

9.1.2 United States Response Rates

The United States response rates, based on the Council of American Survey Research Organizations (CASRO) guidelines, were calculated in accordance with the American Association for Public Opinion Research's (AAPOR) *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*[11] using the assumptions for AAPOR's Response Rate #4.

The United States resolution rate measures the proportion of sampled telephone numbers that could be positively identified as residential or non-residential. This rate was 80.4% for the JCUSH. When called, the majority of the unresolved telephone numbers either rang with no answer or reached persons or machines who "hung up" before identifying themselves. (Most of the remaining unresolved numbers reached answering machines that provided no indication of whether the caller had reached a residence or a business.) This *resolution rate* is one component of the overall response rate.

The second component of the overall response rate is the *cooperation rate*, which measures the proportion of known households within which an interview was completed. The cooperation rate was 62.4% for the JCUSH. The unweighted CASRO response rate was then calculated as the product of the resolution rate (80.4%) and the cooperation rate (62.4%), for an overall United States response rate of 50.2%. Detailed information regarding final sample disposition and United States response rate calculations appear in Table 9.1.2.

Table 9.1.2 Final Sample Disposition

Category	Frequency
Total Out-of-Scope	**17,437**
Total Unresolved	**6,263**
Non-response (known or assumed household)	**3,117**
Completed interview (known household)	**5,183**
Total of selected number	**32,000**

Resolution rate (RR)

$$RR = \frac{\#\ \textit{out-of-scope} + \#\ \textit{non-responding persons} + \#\ \textit{responding persons}}{\textit{Total of selected phone numbers}}\ .$$

[11] The American Association for Public Opinion Research (2004). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 3rd edition. Lenexa, Kansas: AAPOR.

Cooperation rate (CR)

$$CR = \frac{\textit{\# of responding persons}}{\textit{\# of nonresponding persons + \# of responding persons}} .$$

Overall response rate = RR x CR .

Next is an example of how to calculate the overall response rate for the United States using the information found in Table 9.1.2.

$$RR = \frac{(17{,}437 + 3{,}117 + 5{,}183)}{32{,}000} = 0.804 .$$

$$CR = \frac{5{,}183}{3{,}117 + 5{,}183} = 0.624 .$$

Overall response rate = 0.804 x 0.624
= 0.502
= **50.2% .**

9.2 Survey Errors

The estimates derived from this survey are based on a sample of individuals. Hypothetically, somewhat different figures might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. as those actually used. The difference between the estimates obtained from the sample and the results from a complete count under similar conditions is called the *sampling error* of the estimate.

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

9.2.1 Non-sampling Errors

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

A major source of non-sampling errors in surveys is the effect of non-response on the survey results. The extent of non-response varies from partial non-response (failure to answer just one or some questions) to total non-response. Partial non-response to the JCUSH was minimal; once the questionnaire was started, it tended to be completed with very little non-response. Total non-response occurred either because a respondent refused to participate in the survey, or because the interviewer was unable to contact the selected respondent. Total non-response was handled by adjusting the weight of persons who responded to the survey to compensate for those who did not respond. See Section 8 for details of the weight adjustment for non-response.

9.2.2 Sampling Errors

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

For example, suppose hypothetically that one estimates that 20% of respondents aged 18 are smokers and that this estimate is found to have a standard deviation of .007. Then the CV of the estimate is calculated (as a percent) as:

$$(0.007/0.20) \times 100 = 3.5\% \ .$$

Statistics Canada and NCHS commonly use CV results when analyzing data, and they urge users producing estimates from the JCUSH data files to also do so. Please refer to Section 10 for more details on data analysis.

9.2.3 Variance Estimation

In order to determine the quality of the estimate and to calculate the CV, the standard deviation must be calculated. Confidence intervals also require the standard deviation of the estimate. For the JCUSH, it is recommended that the bootstrap method or the Taylor Series Method for variance estimation be used.

9.2.3.1 Bootstrap Method for Variance Estimation

The JCUSH uses a complex survey design, which means that there is no simple formula that can be used to calculate variance estimates. Therefore, an approximate method is needed. The bootstrap method can be used to take into account the sample design information when calculating variance estimates. The bootstrap method, with the use of the Bootvar program provided with the data and discussed in the next subsection, is a method that is fairly easy to use.

The bootstrap method used with the JCUSH data involves the selection of simple random samples known as replicates, and the calculation of the variation in the estimates from replicate to replicate. In each replicate, the survey weight for each record is recalculated. These weights are adjusted and post-stratified according to population estimates information in the same way as the initial weights in order to obtain the final bootstrap weights.

The entire process (selecting simple random samples, recalculating and post-stratifying weights for each stratum) is repeated B times, where B is large. The JCUSH uses B=1,000 to produce 1,000 sets of bootstrap weights, which are provided with the Public Use Microdata File. To obtain a bootstrap variance estimator, the point estimate for each of the B samples must be calculated. The variance of these estimates is the bootstrap variance estimator. A program was developed and can perform all of these calculations for the user: the Bootvar program.

9.2.3.1.1 Bootvar Program for Variance Estimation

The Bootvar program is available in both SAS and SPSS formats. It is made up of macros that compute variances for totals, ratios, differences between ratios and for linear regression and logistic regression. The Bootvar program is provided with the Public Use Microdata File along with bootstrap weights and a document explaining how to modify and use the program to suit users' needs.

9.2.3.2 Taylor Series Method (with SUDAAN) for Variance Estimation

The Taylor series method can be used to estimate variances for totals, ratios, linear regression and logistic regression. The calculation of standard errors for estimates in the JCUSH can be done using statistical software such as SUDAAN.[12] For variance estimation purposes, the JCUSH is treated as a two-stage sample. In SUDAAN applications, WTS_STR and SAMPLEID are used as stratum variables in the SUDAAN NEST statement (household is the PSU, but only one respondent was selected in each household), while WT_SAM is used in the WEIGHT statement. Note that SAMPLEID must first be converted into numeric format (see below).

The following SUDAAN design statements are recommended:

```
sampid = input(sampleid, 12.);

PROC SORT;
    by WTS_STR SAMPID;

PROC...   DESIGN=WR;
NEST      WTS_STR SAMPID;
WEIGHT    WT_SAM;
```

[12] Shah, B.V., Barnwell, B.G., and Bieler, G.S. (1997). *SUDAAN User's Manual, Release 7.5*. Research Triangle Park, NC: Research Triangle Institute.

Subsetted Data Analysis. Frequently, studies using complex survey data are restricted to specific populations subgroups, e.g., persons aged 65 and older. Some users delete all records outside of the domain of interest (e.g., persons aged less than 65 years) in order to work with smaller data files and run computer jobs more quickly. This procedure of keeping only selected records (and list-wise deleting other records) is called subsetting the data. With a subsetted database that is appropriately weighted, correct point estimates (e.g., estimates of population subgroup means) can be produced. However, most software packages that analyze complex survey data using Taylor series linearization incorrectly compute standard errors for estimates calculated for subsetted data. When complex survey data are subsetted, oftentimes the sample design structure is compromised because the complete design information is not available; subsetting data deletes important design information needed for variance estimation. Note that SUDAAN has a SUBPOPN option that allows the targeting of a subpopulation while using the full (unsubsetted) data file that has all sample design information. Using SUDAAN, this section provides two strategies for calculating variances that account for the complex survey design. See a SUDAAN manual for more information.

Strategy 1 Use the MISSUNIT option on the NEST statement with the Method described above for subsetted data:

```
NEST        WTS_STR SAMPID / MISSUNIT;
```

In a WR design with exactly two PSUs per stratum, when some PSUs are removed from the database through the listwise deletion of records outside the population of interest, the MISSUNIT option in SUDAAN "fixes" the estimation to produce standard errors identical to those achieved when using a full dataset with a SUBPOPN statement (see Strategy 2, below). Note that other calculations for design effects, degrees of freedom, and standardization may need to be carried out differently. Users are responsible for verifying the correctness of their results based on subsetted data.

Strategy 2 Use the SUBPOPN statement with the method described above for the full dataset:

```
PROC …          DESIGN=WR;
NEST            WTS_STR SAMPID;
WEIGHT          WT_SAM;
SUBGROUP        (variable names);
LEVELS…         ;
SUBPOPN         SAMPTYPE = 1 and SEX = 2 / NAME="Analysis of Canadian
women";
```

Using the full dataset with the SUBPOPN statement in this example would constrain analysis to Canadian women only (SAMPTYPE = 1 for Canada and SEX = 2 for female) Use of the SUBPOPN statement is equivalent to subsetting the dataset (i.e., deleting the U.S. cases), except that the resulting variance estimates are based on the full design structure for the complete dataset.

10. Guidelines for Tabulation, Analysis and Release

This section outlines the guidelines to be adhered to by users tabulating, analyzing, publishing or otherwise releasing any data derived from the survey microdata file. With 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 NCHS, and, at the same time, they will be able to develop currently unpublished figures in a manner consistent with these established guidelines.

10.1 Rounding Guidelines

In order that estimates for publication or other release derived from this Public Use Microdata Files correspond to those produced by Statistics Canada and NCHS, it is recommended that users adhere to the following guidelines regarding the rounding of such estimates:

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

b) Marginal sub-totals and totals in statistical tables should be derived from their corresponding unrounded components and then should be rounded themselves to the nearest 100 units using normal rounding.

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

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

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

10.2 Sample Weighting Guidelines for Tabulation

The sample design used for the the JCUSH was not self-weighting. That is to say, the sampling weights are not identical for all individuals in the sample. When producing simple estimates, including the production of ordinary statistical tables, users must apply the proper sampling weights.

If proper weights are not used, the estimates derived from the Public Use Microdata Files cannot be considered to be representative of the survey population, and they will not correspond to those produced by Statistics Canada and NCHS.

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

10.2.1 Definitions: Categorical Estimates, Quantitative Estimates

Before discussing how the JCUSH data can be tabulated and analyzed, it is useful to describe the two main types of point estimates of population characteristics that can be generated from the microdata file.

Categorical Estimates

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

Example of Categorical Question:

SMJ1_4 **Do you now smoke cigarettes every day, somedays or not at all?**

1 Everyday
2 Some days
3 Not at all

Quantitative Estimates

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

An example of a quantitative estimate is the average 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, and the denominator is an estimate of the number of individuals who smoke daily.

Example of Quantitative Question:

SMJ1_6 **How many cigarettes do you smoke each day now?**

|_|_| Number of cigarettes

10.2.2 Tabulation of Categorical Estimates

Estimates of the number of people with a certain characteristic can be obtained from the microdata files by summing the final weights of all records possessing the characteristic of interest.

Proportions and ratios of the form $\hat{X} / \hat{Y}$ are obtained by:

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

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

c) dividing the numerator estimate by the denominator estimate.

10.2.3 Tabulation of Quantitative Estimates

Estimates of quantities can be obtained from the microdata files by:

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

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

c) dividing the numerator estimate by the denominator estimate.

For example, to obtain an estimate of the average number of cigarettes smoked each day by individuals who smoke daily, multiply the value of variable **SMJ1_6**[13] by the weight, **WT_SAM**, and then sum this value over those records with a value of "daily" for the variable **SMJ1_4** to obtain the numerator ($\hat{X}$). Sum the final weight of those records with a value of "daily" for the variable SMJ1_4 to obtain the denominator ($\hat{Y}$). Divide ($\hat{X}$) by ($\hat{Y}$) to obtain the average number of cigarettes smoked each day by daily smokers.

[13] See Section 11.2 for variable naming convention.

10.3 Guideline for Statistical Analysis

The JCUSH is based upon a complex design, with stratification and multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents a special situation to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures can differ from what is appropriate in a sample survey framework, with the result that while in many cases the estimates produced by the packages are correct, the variances that are calculated are almost meaningless.

For many analysis techniques (for example linear regression, logistic regression, analysis of variance), a shortcut method exists that can make the application of standard packages more meaningful, with respect to weighting. If the weights on the records are rescaled so that the average weight is one (1), then the results produced by the standard packages will be more reasonable; they still will not take into account the stratification and clustering of the sample's design, but they will take into account the unequal probabilities of selection. The rescaling can be accomplished by dividing each weight in the original analysis by the average of the original weights. Thus, the sum of the original weights is the population size, and the sum of the rescaled weights is the sample size.

10.4 Release Guidelines

Before releasing and/or publishing any estimate from the Public Use Microdata File, users should first determine the number of sampled respondents who contribute to the calculation of the estimate. If this number is less than 10, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. For weighted estimates based on sample sizes of 10 or more, users should determine the coefficient of variation of the estimate and follow the recommended guidelines described in Tables 10.1 and 10.2. Finally, a word of caution: when reporting frequencies, it is important to note that they are underestimates due to item non-response and unknowns.

Table 10.1 Sampling Variability Guidelines Followed by Statistics Canada

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

Table 10.2 Sampling Variability Guidelines Followed by NCHS

Type of Estimate	CV (in %)	Guidelines
Acceptable	$0 \leq CV \leq 30$	Estimates can be considered for general unrestricted release. Requires no special notation.
Marginal	$30 < CV \leq 50$	Such estimates can be released, but it is recommended that they be footnoted with a warning cautioning users that such estimates do not meet the NCHS standard of reliability.
Unacceptable	$CV > 50$	NCHS recommends that such estimates not be released.

11. File Usage

This section starts with a discussion of the ***weight variable.*** This is followed by an explanation of the variable naming conventions that are employed for the JCUSH. The last part of the section discusses alternate approaches to data access available to analysts.

11.1 Use of Weights

Only one weight variable, WT_SAM, appears on the file. This weight variable is applicable to respondents for both countries. ALL VARIABLES ON THE FILE SHOULD BE ANALYZED USING THIS WEIGHT VARIABLE.

(For a more detailed explanation on the creation of this weight variable, see Section 8 of this documentation.)

11.2 Variable Naming Convention

The JCUSH adopted a variable naming convention that allows data users to easily use and identify the data based on module and cycle. The variable naming convention includes the following mandatory requirements: restrict variable names to a maximum of 8 characters for ease of use by analytical software products; identify current and potential future cycles (Cycle 1, 2 …) in the name; and allow conceptually identical variables to be easily identifiable over potential future survey cycles. The variable names for these identical modules and questions should only differ in the cycle position identifying the particular survey occasion on which they were collected.

11.2.1 Variable Name Component Structure in the JCUSH

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

Positions 1-2: Questionnaire section name
Position 3: J: JCUSH Identifier
Position 4: 1: Survey cycle
Position 5: Variable type
Positions 6-8: Question number

For example: The variable from question 3D, Limitation of Activities Module (AHJ1_03D):

Positions 1-2:	**AH**	Limitation of Activities
Positions 3:	**J**	JCUSH identifier
Position 4:	**1**	Cycle 1
Position 5:	**_**	(**_** = collected data)
Positions 6-8:	**03D**	question number and answer option

11.2.2 Positions 1-2: Questionnaire Section Names

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

AD	Administration	AH	Limitation of Activities
DH	Household Contact and Demographics	PS	PAP Smear Test
GH	General Health	MA	Mammography
RA	Restriction of Activities	DE	Dental Visits
CH	Chronic Conditions	IS	Insurance
DP	Depression	RS	Vocational Restriction of Activities
CM	Contact with Mental Health Professionals	SA	Patient Satisfaction
SM	Smoking	PA	Physical Activities
HU	Health Utility Index (HUI)	SD	Socio-demographics Characteristics
HW	Height and Weight	IW	Income and Wealth
HC	Health Care Utilisation	WT	Sample Weights
ME	Use of Medications		

11.2.3 Position 3: The JCUSH Identifier

The naming convention used in the JCUSH is similar to those of Statistics Canada's major health surveys, the Canadian Community Health Survey and the National Population Health Survey (two of the three source surveys for the JCUSH). All variable names in the JCUSH have a J in the third position to differentiate them from questions in the the existing source surveys.

11.2.4 Position 4: Cycle

This is the first cycle of the JCUSH; thus, all variables have a 1 in the fourth position of the name.

11.2.5 Position 5: Variable Type

_	Collected variable	A variable that appeared directly on the questionnaire
D	Derived variable	A variable calculated from one or more collected or coded variables, usually calculated during head office processing (e.g., Health Utility Index)
F	Flag variable	A variable calculated from one or more collected variables (like a derived variable), but usually calculated by the data collection 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)

11.2.6 Positions 6-8: Variable Name

In general, the last three positions follow the variable numbering used on the questionnaire. The letter "Q" used to represent the word "question" is removed, and all question numbers are presented in a two-digit format. For example, question Q01A in the questionnaire becomes simply 01A, and question Q15 becomes simply 15.

For questions that have more than one response option, the final position in the variable naming sequence is represented by a letter. For this type of question, new variables were created to differentiate between a “yes” or “no” answer for each response option. For example, if Q2 had 4 response options, the new variables would be named Q2A for option 1, Q2B for option 2, Q2C for option 3, etc. If only options 2 and 3 were selected, then Q2A = No, Q2B = Yes, Q2C = Yes, and Q2D = No.

11.3 Access to Data

The JCUSH Public Use Microdata File is available for free on the Statistics Canada Web site http://www.statcan.ca/start.html and the National Center for Health Statistics of the United States Centers for Disease Control and Prevention Web site http://www.cdc.gov/nchs/ .

In order to protect the confidentiality of respondents participating in the survey, microdata files must meet stringent security and confidentiality standards required by Canada's *Statistics Act* and *The United States Public Health Service Act* before they are released for public access. 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. Sometimes, these are the variables that are 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 investment.

Statistics Canada and NCHS have procedures in place to prevent the disclosure of confidential data. Information published or otherwise disseminated by Statistics Canada and NCHS is carefully screened to ensure that no confidential data are released. Any attempt to determine the identity of any individual respondent is prohibited. All direct identifiers, as well as any characteristics that might assist in reidentification, are omitted from the public use data files. Any intentional attempt to reidentify the records of respondents violates the assurances of confidentiality given to the providers of the information. By using the Canadian and United States data files, users agree to comply with the following requirements:

1. The data in these files are to be used only for statistical research and data analysis;
2. No attempts will be made to identify any of the records included on the data files; and
3. The data files will not be linked to any other individually identifiable data from other Canadian or United States sources.

Analysts interested in working with United States data that were suppressed to protect confidentiality may access selected unmodified data files at the NCHS Research Data Center (RDC). This facility, designed for the researcher visiting from outside of NCHS, is located at NCHS in Hyattsville, Maryland. Data files housed in the RDC may also be accessed remotely via e-mail. For more information about how to apply for access, analysts may visit:

http://www.cdc.gov/nchs/r&d/rdc.htm.

There is a charge for RDC services.