Statistics Canada Statistique Canada

BRITISH COLUMBIA SMOKING SURVEY HEALTH SURVEY (BCSS) (2006)

MASTER FILE USER GUIDE

December 2008

TABLE OF CONTENTS

1. Introduction

The British Columbia Smoking Survey (BCSS) is a cross-sectional survey that collects information related to the smoking history, mobility history and risk propensity of British Columbia residents. BCSS is a follow-up survey based on the Canadian Community Health Survey (CCHS) Cycle 3.1 (2005).

The BCSS collected responses from:

- Persons aged 18 or older living in private occupied dwellings
- Non-proxy respondents living in British Columbia at the time of their interview who agreed to link their CCHS 3.1 data.

The CCHS excluded individuals living on Indian Reserves and on Crown Lands, institutional residents, full-time members of the Canadian Armed Forces and residents of certain remote regions. BCSS was conducted from February 2006 to May 2006.

The purpose of this document is to facilitate the manipulation of the BCSS master file and to describe the methodology used.

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

For custom tabulations or general data support:
Client Custom Services, Health Statistics Division: 613-951-1746
E-mail: hd-ds@statcan.ca

2. Background

The central objective of the British Columbia Smoking Survey (BCSS) was to gather information related to the smoking history, mobility history and risk propensity of British Columbia residents. The survey was sponsored by the B.C. Ministry of Health. The BCSS was a cross-sectional survey that was a follow-up to the Canadian Community Health Survey, Cycle 3.1 conducted in 2005.

2.1 The Data Model

Respondents were asked their smoking history from the time they began smoking until the present (time of the survey) or until they stopped smoking.

The first question in the BCS module asked if the respondent had smoked a total of 100 or more cigarettes (about 4 packs) in their lifetime (BCSF_100). If the respondent answered yes, they were then asked a series of questions about the following:

- The age they started smoking at least one cigarette a month (BCSF_102).
- How many days in a month they smoked at least one cigarette (BCSF_103).
- How many cigarettes they usually smoked each day (BCSF_104).

After this first series of questions, the respondent was asked if they ever had a period of at least one year when they stopped smoking, changed the number of cigarettes smoked, or changed the number of days in a typical month they smoked. If the respondent never had any change in smoking habits (answered no to BCSF_107), they did not continue in this section of the questionnaire. If they answered that they had a change in smoking habits (answered yes to BCSF_107), they were asked a series of questions up to nine times which reflected the changes in their habits over a time period. Each of the changes had to have lasted for at least one year. The nine possible iterations of this section asked the following:

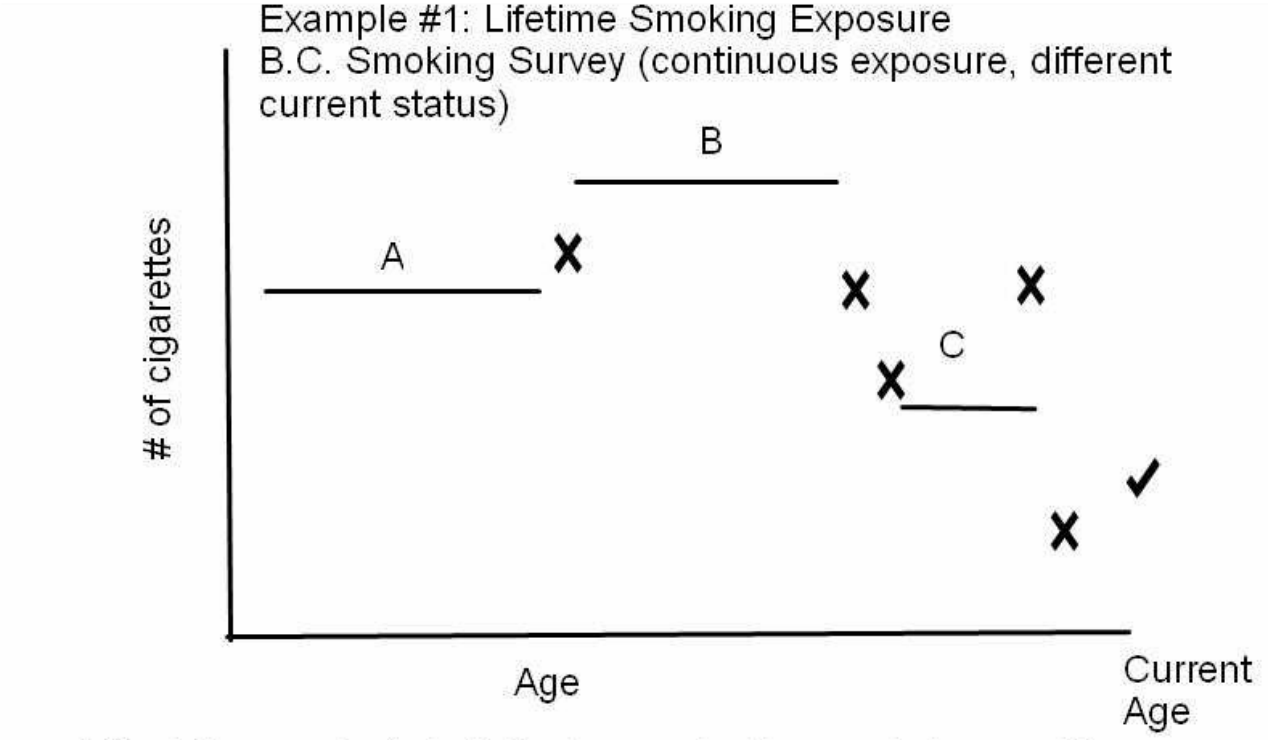
- Was there was a period of at least one year when they stopped smoking, changed the number of cigarettes smoked, or changed the number of days in a typical month smoked (BCSF_107).
- Age when their pattern changed (BCSF_108).
- How many days in a month they smoked at least one cigarette (BCSF_110).
- Number of cigarettes they usually smoked on each day (BCSF_111).

In the three examples given below, each letter represents a different period in the respondent's smoking history.

- The letter A represents the first series of questions in the BCS module. The respondent smoked 100 or more cigarettes in their lifetime (BCSF_100 = YES), and gave a response indicating the age they started smoking and the number of cigarettes they usually smoked on a specified number of days per month for a time period of at least one year (BCSF_102, BCSF_103, BCSF_104).

- The letters B and C represent periods where a change in smoking pattern occurred. If the respondent answered YES to BCSF_107, there was a period of at least one year when they stopped smoking, changed the number of cigarettes smoked, or changed the number of days in a typical month smoked. They were then asked about the age at which this change occurred and the number of cigarettes they usually smoked on a specified number of days per month for a time period of at least one year (BCSF_108, BCSF_110, BCSF_111).

- The letter X is used to represent periods of time that were less then one year in length where a change occurred. Information was not collected for these periods.



In Example #1, A, B and C represent periods of at least a year when the respondent was smoking a specified number of cigarettes for a specified number of days per month. The questions were asked for each time period after the initial response (A) to a maximum of nine periods (B, C, etc). In this example, the respondent provided information for 3 time periods that reflected a change in smoking patterns that lasted for at least one year. The Xs represent periods of less than a year when the consumption of cigarettes was more variable.

In this example, the letter A represents the following information:

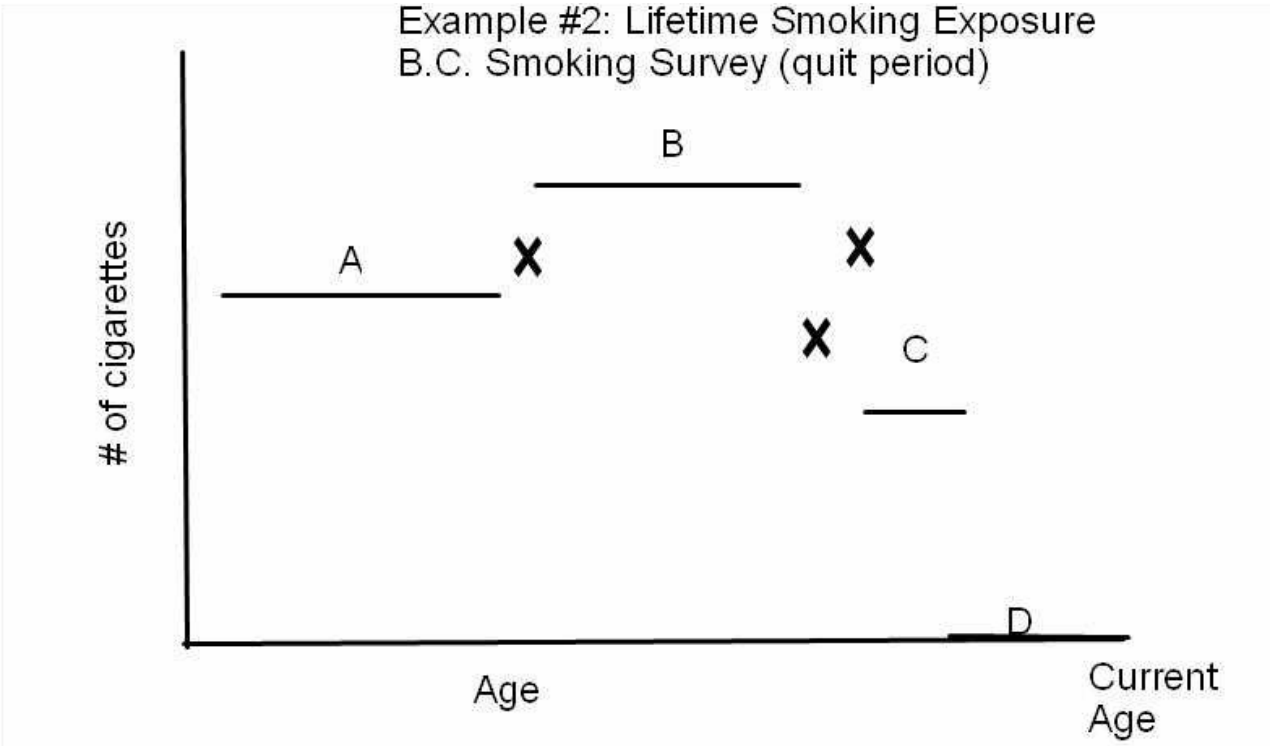
- The age they started smoking at least one cigarette a month (BCSF_102) and the number of cigarettes they usually smoked each day (BCSF_104).

The letters B and C represent the following information:

- A period of at least one year when they stopped smoking, changed the number of cigarettes smoked, or changed the number of days in a typical month smoked (BCSF_107 = YES). For each period of change in their smoking pattern (B and C) the age when their pattern changed (BCSF_108), and the number of cigarettes they usually smoked on each day (BCSF_111) are illustrated above.

The areas in the model designated with an X represent a change in smoking pattern that lasted less than one year.

The number of cigarettes smoked at the time of the interview is denoted by a checkmark.



In Example #2, A, B and C represent periods of at least a year when the respondent was smoking a specified number of cigarettes for a specified number of days per month. D represents a period of at least one year when the respondent had quit smoking (at the time of the interview, they still were not smoking). The questions were asked for each time period after the initial response (A) to a maximum of nine periods (B, C, etc). In this example, the respondent provided information for four time periods that reflected a change in smoking patterns that lasted for at least one year. The Xs represent periods of less than a year when the consumption of cigarettes was more variable.

In this example, the letter A represents the following information:

- The age they started smoking at least one cigarette a month (BCSF_102) and the number of
  cigarettes they usually smoked each day (BCSF_104).

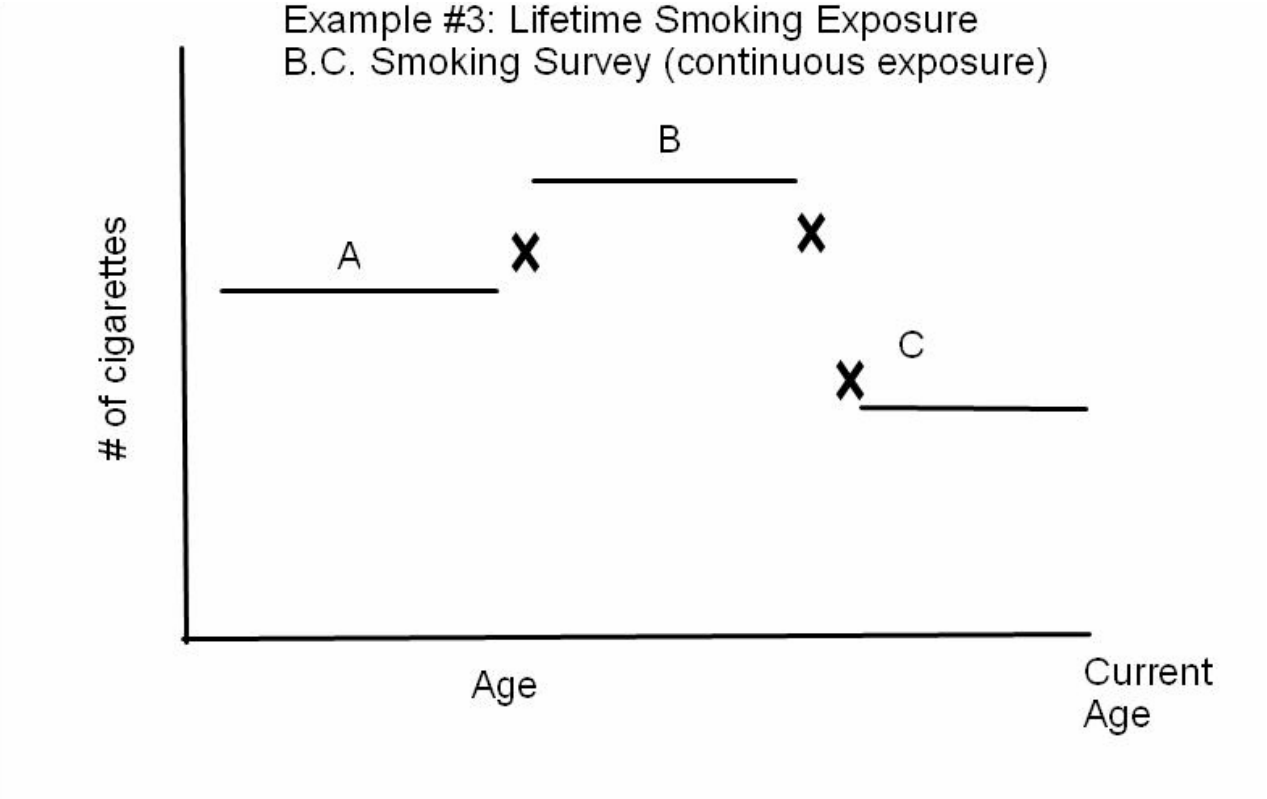
The letters B and C represent the following information:

- A period of at least one year when they stopped smoking, changed the number of cigarettes smoked, or changed the number of days in a typical month smoked (BCSF_107 = YES). For each period of change in their smoking pattern (B and C) the age when their pattern changed (BCSF_108) and the number of cigarettes they usually smoked on each day (BCSF_111) are illustrated above.

The letter D represents the following information:

- The period of time when the respondent had quit smoking for at least one year (BCSF_107).

The area in the model designated with an X represents a change in smoking pattern that lasted less than one year.



In Example #3, A, B and C represent periods of at least a year when the respondent was smoking a specified number of cigarettes for a specified number of days per month. The questions were asked for each time period after the initial response (A) to maximum of nine periods (B, C, etc). In this example, the respondent provided information for three time periods that reflected a change in smoking patterns that lasted for at least one year. The Xs represent periods of less than a year when the consumption of cigarettes was more variable.

In this example, the letter A represents the following information:

- The age they started smoking at least one cigarette a month (BCSF_102) and the number of cigarettes they usually smoked each day (BCSF_104).

The responses in B and C represent the following information:

- A period of at least one year when they stopped smoking, changed the number of cigarettes smoked, or changed the number of days in a typical month smoked (BCSF_107 = YES). For each period of change in their smoking pattern (B and C) the age when their pattern changed (BCSF_108) and the number of cigarettes they usually smoked on each day (BCSF_111) are illustrated above.

The area in the model designated with an X represents a change in smoking pattern that lasted less than one year.

3. Objective

The central objective of the British Columbia Smoking Survey (BCSS) was to gather information related to the smoking history, mobility history and risk propensity of British Columbia residents.

4. Survey content

This section provides a general discussion of the consultation process used in survey content development and gives a summary of the final content selected for inclusion in BCSS.

The content was developed by the Health Statistics Division based on the objectives and requirements specified by the client. The client was consulted throughout the development and testing of the questionnaire.

4.1 Qualitative Testing

Two qualitative tests were conducted to assess the content and flow of the BCSS questionnaire. The first was conducted in November 2004 in Ottawa, Ontario. Approximately 25 respondents, covering a cross-section of current and former smokers, including daily and occasional, were represented in face-to-face interviews. The frame used to select respondents for the interviews was the CCHS Cycle 2.1 (2003). All qualitative interviews were conducted by representatives of the Questionnaire Design Resource Centre (QDRC) and observed by the Health Statistics Division BCSS project team.

A key finding from the November test was that respondents had difficulty recalling their various smoking patterns over time. In order to facilitate collecting this history, it was determined that establishing a linear timeline within the questionnaire would aid respondents in remembering their periods of smoking. For example, at the beginning of the section to collect the smoking history, respondents were asked when they had smoked their first cigarette and then informed questions would follow to determine successive periods of smoking or quitting (i.e., the linear timeline).

A second and final qualitative test was conducted in February 2005 in Vancouver BC. Approximately 25 respondents, covering a cross section of current and former smokers, including daily and occasional smokers were represented in face-to-face interviews. The frame used to select respondents for the interviews was the CCHS Cycle 2.1 (2003). All qualitative interviews were conducted by representatives of the Questionnaire Design Resource Centre (QDRC) and observed by the Health Statistics Division BCSS project team.

The key finding was that the timeline approach proved to be a success and most respondents found it a very useful in helping them to gather their thoughts about when and how much they had smoked over various periods of time in their lives.

4.2 Field Test

In November 2005, a pilot test of the BCSS was conducted. The frame used to select approximately 300 respondents for the interviews was the CCHS Cycle 2.1 (2003). The main objective was to determine whether the content developed and successfully tested during the qualitative testing (face-to-face interviews) the previous February could be successfully administered by telephone interviews. Data collected during the field test indicated both respondents and interviewers felt BCSS was suitable for and could be successfully administered through telephone interviewing.

4.3 Final Questionnaire Content

This section outlines the modules within the questionnaire component (C2) of BCSS and details how these modules are relevant to the survey. The BCSS C2 was divided into four modules; each module collected a specific type of information:

Module	Content
BCS	**British Columbia Smoking History/Lifetime Exposure to Cigarettes**: This module collected information on whether or not respondents smoked 100+ cigarettes in their lifetime, a timeline of amount, frequency as well as the brands smoked. The module was asked of all respondents.
BCM	**Mobility**: This module collected information on time respondents spent living inside and outside B.C. The module was only asked of respondents who were former and/or current smokers.
BCD	**Driving and Safety:** This module collected information on the driving habits and driving safety of both drivers and passengers in a motor vehicle. This module was asked of all respondents.
BCA	**Administration:** This module collected administrative information including permission to link information with the previous CCHS and administrative health records from the province, valid health card information and language of interview. This module was asked of all respondents.

5. Sample Design

5.1 Target population

The target population consists of B.C. residents aged 18 and over, living in private occupied dwellings at the time of the CCHS Cycle 3.1 interview in 2005. Individuals living on Indian Reserves and on Crown Lands, institutional residents, full-time members of the Canadian Armed Forces and residents of certain remote regions are excluded.

5.2 Sample frame and sampling strategy

The survey frame for the BCSS sample was created by using the respondents from the Canadian Community Health Survey (CCHS) Cycle 3.1 (2005). Since the CCHS data would be linked to the BCSS data, only those individuals who agreed to link their CCHS Cycle 3.1 data were included on the BCSS frame. Also, only those CCHS respondents aged 18 years or older and living in BC at the time of the Cycle 3.1 interview were contacted for the BCSS. All proxy CCHS Cycle 3.1 respondents were not included on the BCSS frame because of the personal nature and recall demands of the BCSS questionnaire. For CCHS, proxy interviews are allowed in cases where the respondent does not have the capacity to complete the interview because of physical or mental limitations. Finally, any partially completed CCHS Cycle 3.1 interviews were excluded from the BCSS frame. The reason for this exclusion was two-fold. First, the question concerning agreement to link the CCHS Cycle 3.1 data was asked in the last module of the questionnaire and most partially complete interviews did not reach this point. Second, since the CCHS data would be linked with the BCSS data, it was not desirable to have large portions of missing data. All units that make up the BCSS frame are included in the BCSS sample.

Since this was a targeted respondent survey, only the particular member of a household who had completed a CCHS Cycle 3.1 interview was eligible to be followed up for the BCSS.

5.3 Sample allocation over the collection period

The CCHS Cycle 3.1 sample in B.C. came from 16 health regions (HR) covering the entire province and was allocated proportionally to the square root of the population size of the HR. There were between 500 and 1,600 respondents in each HR. The distribution of the CCHS sample by gender and age was very similar to the actual distribution in the province, with the exception of a slight over-representation of individuals aged 65 and over.

For the BCSS, the collection period for the survey was from February 2006 to May 2006. During this time, there were two samples sent for collection. A sample of approximately 5,900 respondents was assigned at the end of January 2006, representing most eligible respondents who had a CCHS Cycle 3.1 interview from January 2005 to June 2005. A second sample of approximately 5,700 respondents was delivered at the end of March 2006, comprising all other eligible CCHS Cycle 3.1 respondents. A total of 11,587 records were delivered for possible BCSS interviews.

Contact information was not available for all CCHS Cycle 3.1 respondents. Some respondents of the area frame sample did not have or did not provide a valid telephone number. Since the BCSS was conducted over the phone, all respondents for which no phone number was available were identified for tracing. Different sources were available and used to try to find telephone numbers and to contact the respondents that were selected for BCSS. Note that tracing was not only done for the respondents without a phone number, but also for the ones who had moved between the time of their CCHS and BCSS interviews. In addition to those problematic cases, some respondents were not willing to give their real name during the CCHS interview. For those respondents, a fictitious name was used to identify them in CCHS. In order to help the interviewers find the person they were looking for, a complete listing of all respondents with fictitious names was provided, including all members of their household. The following variables were provided for each member of those households: age, relationship to respondent, sex and marital status. These special cases were assigned to experienced interviewers given that they were more difficult to do deal with.

6. Data collection

6.1 Computer-assisted interviewing

Collection for BCSS took place from February to May 2006. Over the collection period, a total of 9,316 valid interviews were conducted using computer assisted telephone interviewing (CATI).

Computer-assisted interviewing (CAI) offers two main advantages over other collection methods. First, CAI offers a Case Management System and data transmission functionality. This case management system automatically records important management information for each attempt on a case and provides reports for the management of the collection process. CAI also provides an automated call scheduler, i.e. a central system to optimize the timing of call-backs and the scheduling of appointments used to support CATI collection.

The case management system routes the questionnaire applications and sample files from Statistics Canada's main office to regional collection offices (in the case of CATI). Data returning to the main office takes the reverse route. To ensure confidentiality, the data is encrypted before transmission. The data are then unencrypted when they are on a separate secure computer with no remote access.

Second, CAI allows for custom interviews for every respondent based on their individual characteristics and survey responses. This includes:

- Questions that are not applicable to the respondent are skipped automatically.
- 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.

- 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.

6.2 BCSS application development

For BCSS, a computer-assisted telephone interview (CATI) application was utilized. At the time of development, a newly standardized entry and exit application components/procedures for targeted respondents had been developed at Statistics Canada. Some customization of the entry and exit components was allowed for BCSS. The application consisted of entry, survey content (known as the C2), and exit components.

Entry and exit components contain standard sets of questions designed to guide the interviewer through contact initiation, respondent confirmation, tracing (if necessary) and determination of case status. The C2 consists of the BCSS modules themselves which made up the bulk of the application. The application and survey content was pilot tested in November 2005.

Feedback from the pilot test was used to make modification/improvements to the CATI application. Once these modifications were complete, final testing of the full application began. This consisted of three stages of internal testing: block, integrated and end to end.

Block level testing consists of independently testing each content module or "block" to ensure skip patterns, logic flows and text, in both official languages, are specified correctly. Skip patterns or logic flows across modules are not tested at this stage as each module is treated as a stand alone questionnaire. Once all blocks are verified by several testers, they are added together along with entry and exit components into an integrated application. This newly integrated application is then ready for the next stage of testing.

Integrated testing occurs when all of the tested modules are added together, along with the entry and exit components, into an integrated application. This second stage of testing ensures that key information such as age and gender are passed from the sample file to the C2 and exit components of the applications. It also ensures that variables affecting skip patterns and logic flows are correctly passed between modules within the C2. Since, at this stage, the applications essentially function as they will in the field; all possible scenarios faced by interviewers are simulated to ensure proper functionality. These scenarios test various aspects of the entry and exit components including; establishing contact, confirming that the correct respondent has been found, determining whether a case is in scope and creating appointments.

End to end testing occurs when the fully integrated application is placed in a simulated collection environment. The application is loaded onto computers that are connected to a test server. Data is then collected, transmitted and extracted in real time, exactly as it would be done in the field. This last stage of testing allows for the testing of all technical aspects of data input, transmission and extraction for the BCSS application. It also provided a final chance of finding errors within the entry, C2 and exit components.

6.3 Interviewer training

A team of Statistics Canada headquarters representatives went to Edmonton in February 2006 to train the Regional Office project manager and a small team of interviewers for the survey. Members of the survey team attended these training sessions to offer additional support and clarify any questions or concerns that may have arisen.

The focus of these sessions were to get interviewers comfortable using the BCSS application and familiarise interviewers with survey content. The training sessions focused on:

- goals and objectives of the survey
- survey methodology
- application functionality
- review of the questionnaire content and exercises
- use of mock interviews to simulate difficult situations and practise potential non-response situations
- survey management

One of the key aspects of the training was a focus on minimizing non-response. Exercises to minimise non-response were prepared for interviewers. The purpose of these exercises was to have the interviewers practice convincing reluctant respondents to participate in the survey.

6.4 The interview

Sample units selected from the frame were interviewed from a centralised call centre using CATI. The CATI interviewers were supervised by a senior interviewer located in the same call centre.

To ensure the quality of the data collected, interviewers were instructed to make every effort to conduct the interview with the selected respondent in privacy. In situations where this was unavoidable, the respondent was interviewed and flags on the Master File indicate whether the interviewer determined that somebody other than the respondent was present during the interview (BCAF_N10) and whether the interviewer felt that the respondent's answers were influenced by the presence of the other person (BCAF_N11).

To ensure the best possible response rate attainable, many practices were used to minimise non-response, including:

Introductory letters

Before the start of each of the two collection periods, introductory letters explaining the purpose of the survey were sent to the sampled households. These explained the importance of the survey and provided examples of how BCSS data would be used.

Initiating contact

Interviewers were instructed to make all reasonable attempts to obtain interviews. When the timing of the interviewer's call was inconvenient, an appointment was made to call back at a more convenient time. Numerous call-backs were made at different times on different days.

Refusal conversion

For individuals who at first refused to participate in the survey, a letter was sent from the Statistics Canada Regional Office to the respondent, stressing the importance of the survey and the household's collaboration. This was followed by a second call from a senior interviewer, a project supervisor or another interviewer to try to convince the respondent of the importance of participating in the survey.

Language barriers

To remove language as a barrier to conducting interviews, the Statistics Canada Regional Office recruited interviewers with a wide range of language competencies. When necessary, cases were transferred to an interviewer with the language competency needed to complete an interview. Chinese and Punjabi were the most common language barriers identified by the Edmonton regional office.

Proxy interviews

Proxy interviews were not permitted.

6.5 Field operations

The BCSS sample was divided into two two-month overlapping collection periods. The regional collection office was instructed to use the first 4 weeks of each collection period to resolve the majority of the sample, with next 4 weeks being used to finalize the remaining sample and to follow up on outstanding non-response cases. All cases were to have been attempted by the second week of each collection period.

Transmission of cases from the Edmonton regional office to head office was the responsibility of the regional office project supervisor, senior interviewer and the technical support team. These transmissions were performed nightly and all completed cases were sent to Statistics Canada’s head office.

6.6 Quality control and collection management

During the BCSS collection cycle, several methods were used to ensure data quality and to optimize collection. These included using internal measures to verify interviewer performance and the use of a series of ongoing reports to monitor various collection targets and data quality.

CATI interviewers were randomly chosen for validation. Validation during CATI collection consisted of senior interviewers monitoring interviews to ensure proper techniques and procedures (reading the questions as worded in the applications, not prompting respondents for answers, etc.) were followed by the interviewer.

A series of reports were produced to effectively track and manage collection targets and to assist in identifying other collection issues.

Cumulative reports were generated at the end of each collection period, showing response, and link rates. At the end of data collection, a response rate of 80% was achieved. A total of 87% of the BCSS respondents agreed to link while 72% of the BCSS respondents agreed to link and reported a valid Health Number. Customised reports were also created and used to examine specific data quality issues that arose during collection.

7. Data processing

7.1 Editing

Most editing of the data was performed at the time of the interview by the computer-assisted interviewing (CAI) application. It was not possible for interviewers to enter out-of-range values and flow errors were controlled through programmed skip patterns. For example, CAI ensured that questions that did not apply to the respondent were not asked.

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 Head Office. Inconsistencies were usually corrected by setting one or both of the variables in question to "not stated".

7.2 Coding

Pre-coded answer categories were supplied for all suitable variables. Interviewers were trained to assign the respondent’s answers to the appropriate category.

7.3 Creation of derived and grouped variables

To facilitate data analysis and to minimise the risk of error, two variables on the file have been derived using items found on the BCSS questionnaire. Derived variables generally have a "D" in the fifth character of the variable name. In both cases one or more variables were used in combination to create a new variable. For more information on the naming convention, please go to Section 12.3.

7.4 Weighting

The principle behind estimation in a probability sample such as the BCSS is that each person in the sample "represents", besides himself or herself, several other persons not in the sample. For

example, in a simple random 2% sample of the population, each person in the sample represents 50 persons in the population. In the terminology used here, it can be said that each person has a weight of 50.

The weighting phase is a step that calculates, for each person, his or her associated sampling weight. This weight appears on the data file and must be used to derive meaningful estimates from the survey. For example, if the number of individuals who smoke daily is to be estimated, it is done by selecting the records referring to those individuals in the sample having that characteristic and summing the weights entered on those records.

Details of the method used to calculate sampling weights are presented in Section 8.

8. Weighting

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

As described in Section 5, the BCSS survey frame is based on respondents to the CCHS Cycle 3.1 (2005). The starting point for the BCSS weighting process is the pre-post-stratified CCHS Master weight. For more information on this weight see the CCHS Cycle 3.1 user guide.

8.1 Sample weighting

Table 8.1 presents an overview of the different adjustments that are part of the weighting strategy for BCSS, in the order in which they are applied.

Table 8.1: Weighting Steps for BCSS

BC1 – Selection Criteria Adjustment
BC2 – Post-stratification of CCHS weights using BCSS age groups
BC3 – Non-response in BCSS
BC4 – Final Post-stratification

8.1.2 Selection Criteria Adjustment

Since BCSS is a supplementary survey based on the CCHS Cycle 3.1 (2005), the starting point for the BCSS weighting process was the pre-post-stratified CCHS master weight. The first step was to remove all CCHS units from outside the province of BC from the CCHS weight file, since these units were not representative of anyone in the target population for BCSS. Given that the CCHS design was stratified within each province, and that each CCHS adjustment was carried out within the province, the CCHS weights outside of BC were independent of those within BC, thus allowing for the simple removal of these units from the file.

The next step was to remove all of the CCHS proxy interviews and partially complete interviews, as well as the respondents who did not agree to link their CCHS data. This group of CCHS respondents which were excluded from the BCSS frame had an initial CCHS weight representing other people living in BC. Therefore, their weights had to be redistributed to the remaining CCHS respondents. This was done by creating response homogeneous groups (RHG) using the software KnowledgeSeeker.[1]

The adjustment was calculated within each RHG as follows:

$$\text{adjBC1} = \frac{\text{Sum of CCHS pre - post - stratified master weight for all units}}{\text{Sum of weight for all non - proxy, complete interviews with agreement to link CCHS data}}$$

The weight wgtBC1 was calculated as wgti2b*adjBC1, where wgti2b is the CCHS pre-post-stratified master weight. After the adjustment was calculated, the excluded units were dropped from the file and the adjustment was applied to the remaining units.

8.1.3 Post-Stratification of CCHS weights using BCSS age groups

At this point, the only units left on the data file that were not included on the BCSS frame were the CCHS respondents under the age of 18. Before these units were removed from the file, the data were post-stratified using the 2005 provincial projections for 2 age groups, namely 12-17 and 18+. The population projections used were based on the 2001 Census. The average of the 2005 monthly estimates for each age group was used as a known total. The adjustment was computed within each post-strata as follows:

$$adjBC2 = \frac{\text{Population estimate for the province : age group of the respondent}}{\text{Sum of wgtBC1 for the province : age group of the respondent}}$$

The weight wgtBC2 was calculated as wgtBC1*adjBC2. After the post-stratification, the CCHS respondents under the age of 18 were dropped from the file. Thus, the units remaining on the file represented the entire BCSS sample, and the weights were representative of the BCSS target population.

8.1.4 Non-response in BCSS

Although all CCHS respondents were targeted for the BCSS, not all individuals were respondents for the BCSS. In this step, the weights of non-responding units in BCSS were distributed amongst responding units. This included the redistribution of the CCHS duplicate cases, which were treated as non-respondents in BCSS. The BCSS sample file contained cases which were known duplicates on CCHS. Duplicates occurred in CCHS when a respondent was chosen on both the area and the phone frame, and chose to respond to the survey twice. They often gave different answers and the interviews were completed at different times throughout the year. Therefore, both

[1] Angoss Software, 2005

versions were kept on the CCHS Cycle 3.1 and given separate weights. These duplicate CCHS cases were discovered using personal information such as name, age, date of birth, phone number and health insurance number. In order to minimize the response burden, these people were only contacted once for BCSS and their duplicate record was treated as non-response. There were no partially complete interviews in BCSS, so only units with fully complete interviews were considered respondents.

For this step, non-response was modeled through chi-squared tests using the software KnowledgeSeeker. RHGs were created using auxiliary variables collected in both CCHS and BCSS. This adjustment was calculated within each RHG as follows:

$$\text{adjBC3} = \frac{\text{Sum of wgtBC2 for all units}}{\text{Sum of wgtBC2 for all responding units}}$$

This adjustment was applied to wgtBC2 to obtain wgtBC3. After the adjustment was calculated, the non-responding units (and the CCHS duplicates) were dropped from the file and the adjustment was applied to the remaining units.

8.1.5 Post-stratification

The final step in obtaining the BCSS master weight was to post-stratify to projected 2005 population counts at the provincial level. Post-stratification was done to ensure that the sum of the final weights corresponded to the population projections defined at the province level for 8 age-sex groups, namely 18-29, 30-44, 45-64 and 65+. The population projections were based on 2001 Census counts. The rounded average of the 2005 monthly population estimates for each age-sex group was used as a known total. The adjustment was computed within each post-strata as follows:

$$adjBC4 = \frac{\text{Population estimate for the province : age - sex group of the respondent}}{\text{Sum of wgtBC3 for the province : age - sex group of the respondent}}$$

The final weight was calculated as wgtBC3*adjBC4. Consequently, the weight wgtBC4 corresponds to the ***final BCSS weight*** that can be found on the data file with the variable name WTSE_M.

8.2 Bootstrap Weights

Coordinated bootstrap weights were used for BCSS because of its dependence on the CCHS sample. Hence, the starting point for the BCSS bootstrap weights was the 500 replicates from the pre-post-stratified CCHS master bootstrap file. Each bootstrap replicate was adjusted using the four adjustments listed in table 8.1.

9. Data quality

9.1 Response rates

At the end of data collection, a response rate of 80% was achieved. A total of 87% of the BCSS respondents agreed to link while 72% of the BCSS respondents agreed to link and reported a valid Health Number.

9.2 Survey Errors

The estimates derived from this survey are based on a sample of individuals. Somewhat different figures might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. than 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 which 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 BCSS. Quality assurance measures were implemented at each step of data collection and processing to monitor the quality of the data. These measures included the use of highly skilled interviewers, extensive training with respect to the survey procedures and questionnaire, and the observation of interviewers to detect problems. Testing of the CAI application and field tests were also essential procedures to ensure that data collection errors were minimized.

A major source of non-sampling errors in surveys is the effect of ***non-response*** on the survey results. The extent of non-response varies from partial non-response (failure to answer just one or some questions) to total non-response. Partial non-response to the BCSS was minimal; once the questionnaire was started, it tended to be completed with very little non-response. Total non-response occurred either because a person refused to participate in the survey or because the interviewer was unable to contact the selected person. 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 on 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.

For example, suppose hypothetically that it is estimated that 25% of British Columbia residents aged 18 and over are regular smokers and that this estimate is found to have a standard deviation of 0.003. Then the CV of the estimate is calculated as:

$$(0.003/0.25) \times 100\% = 1.20\%$$

Statistics Canada commonly uses CV results when analyzing data and urges users producing estimates from the BCSS data files to also do so. For details on how to determine CVs, see Section 11. For guidelines on how to interpret CV results, see the table at the end of Sub-section 10.4.

10. Guidelines for tabulation, analysis and release

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

10.1 Rounding guidelines

In order that estimates for publication or other release derived from this data file correspond to those produced by Statistics Canada, users are urged to adhere to the following guidelines regarding the rounding of such estimates:

a) Estimates in the main body of a statistical table are to 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 proceeding digit is incremented by 1;

b) Marginal sub-totals 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 normal rounding;

c) Averages, proportions, rates and percentages are to be computed from unrounded components (i.e., numerators and/or denominators) and then are to 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) 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 normal rounding;

e) In instances where, due to technical or other limitations, a rounding technique other than normal rounding is used resulting in estimates to be published or otherwise released that differ from corresponding estimates published by Statistics Canada, users are 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.

10.2 Sample weighting guidelines for tabulation

The sample design used for this survey 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 weight. If proper weights are not used, the estimates derived from the data file cannot be considered to be 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.

10.2.1 Definitions: categorical estimates, quantitative estimates

Before discussing how the survey 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 data 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. How often individuals fasten their seat belts when driving is an example of such an estimate. An estimate of the number of persons possessing a certain characteristic or exhibiting certain behaviours may also be referred to as an estimate of an aggregate.

Example of categorical question:

> *How often do you fasten your seat belt when you drive a motor vehicle? **(BCDF_02)***
>
> __ Always
> __ Most of the time
> __ Rarely
> __ Never

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 its denominator is an estimate of the number of individuals who smoke daily.

Example of quantitative question:

> *About how many cigarettes do you usually smoke each day? **(BCSF_136)***
> |_|_| 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 data file 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 ($\hat{Y}$); then

c) dividing the numerator estimate by the denominator estimate.

10.2.3 Tabulation of quantitative estimates

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

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

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

c) dividing the numerator estimate by the denominator estimate.

For example, to obtain the estimate of the average number of cigarettes smoked each day when an individual started smoking for individuals who started smoking on a daily basis from the beginning, first compute the numerator ($\hat{\boldsymbol{X}}$) by summing the product between the value of variable BCSF_104 and the weight WTSF_M. Next, sum this value over those records with a value of "31" to the variable BCSF_103. The denominator ($\hat{\boldsymbol{Y}}$) is obtained by summing the final weight of those records with a value of "31" to the variable BCSF_103. Divide ($\hat{\boldsymbol{X}}$) by ($\hat{\boldsymbol{Y}}$) to obtain the average number of cigarettes smoked each day, when an individual started smoking, for individuals who were daily smokers when they started smoking.

In this example, the variables BCSF_103 and BCSF_104 contain the answers provided by the respondent to the following questions:

> *When you were [age when the respondent started smoking at least one cigarette a month] years old, how many days in a typical month did you smoke at least one cigarette? (**BCSF_103**)*
>
> |_|_| Number of days (if the respondent was smoking every day, 31 is entered)
>
> *When you were [age when the respondent started smoking at least one cigarette a month] years old, how many cigarettes did you usually smoke each day? (**BCSF_104**)*
>
> |_|_| Number of cigarettes

10.3 Guidelines for statistical analysis

The BCSS 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 problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures 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 method exists that can make the application of standard packages more meaningful. If the weights on the records are rescaled so that the average weight is one (1), then the results produced by the standard packages will be more reasonable; they still will not take into account the stratification and clustering of the sample's design, but they will take into account the unequal probabilities of selection. The rescaling can be accomplished by using in the analysis a weight equal to the original weight divided by the average of the original weights for the sampled units (people) contributing to the estimator in question.

10.4 Release guidelines

Before releasing and/or publishing any estimate from the data file, users must first determine the number of sampled respondents having the characteristic of interest (for example, the number of respondents who smoke when interested in the proportion of smokers for a given population). If this number is less than 30, the un-weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. For weighted estimates based on sample sizes of 30 or more, users should determine the coefficient of variation of the rounded estimate and follow the guidelines below.

Table 10.1 Sampling variability guidelines

Type of Estimate	CV (in %)	Guidelines
Acceptable	$0.0 \leq CV \leq 16.6$	Estimates can be considered for general unrestricted release. Requires no special notation.
Marginal	$16.6 < CV \leq 33.3$	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."

11. Variances and coefficients of variation

The computation of exact coefficients of variation is not a straightforward task since there is no simple mathematical formula that would account for all BCSS sampling frame and weighting aspects. Therefore, other methods such as re-sampling methods must be used in order to estimate measures of precision. Among these methods, the bootstrap method is the one recommended for analysis of BCSS data.

The computation of coefficients of variation (or any other measure of precision) with the use of the bootstrap method requires access to information that is considered confidential. This computation must be done using the Master file. Access to the Master file is discussed in section 12.

For the computation of coefficients of variation, the bootstrap method is advised. A macro program, called "Bootvar", was developed in order to give users easy access to the bootstrap method. The Bootvar program is available in SAS and SPSS formats, and is made up of macros that calculate the variances of totals, ratios, differences between ratios, and linear and logistic regressions.

Although some standard statistical packages allow sampling weights to be incorporated in the analyses, the variances that are produced often do not take into account the stratified and clustered nature of the design properly, whereas the exact variance program would do so.

12. File usage

This section begins by describing the master file, the weight variable of the master file and an explanation of how it should be used when doing tabulations. This is followed by an explanation of the variable naming convention that is employed by the BCSS and CCHS. Finally the topic of custom tabulations and contacts for this are given.

12.1 Master file

The master file contains all variables and all records from the survey collected during a collection period. These files are accessible at Statistics Canada for internal use and are also subject to custom tabulation requests.

12.2 Use of weight variable

The weight variable **WTSF_M** represents the BCSS sampling weight. For a given respondent, the sampling weight can be interpreted as the number of people the respondent represents in the population. This weight must always be used when computing statistical estimates in order to make inference at the population level possible. The production of un-weighted estimates is not recommended. The sample allocation, as well as the survey design specifics can cause such results to not correctly represent the population. Refer to section 8 on weighting for a more detailed explanation on the creation of this weight.

12.3 Variable naming convention

The BCSS adopted a variable naming convention that allows data users to easily use and identify the data based on module and cycle. The convention follows a similar standard of the CCHS as it is a supplement to the CCHS 3.1. The variable naming convention follows the mandatory requirement of restricting variable names to a maximum of 8 characters for ease of use by analytical software products.

12.3.1 Variable name component structure in BCSS

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

Positions 1-2:	Reference to BC Smoking Survey (BC)
Position 3:	Module reference (S – Smoking, M – Mobility, D – Driving and Safety and A – Administration)
Position 4:	Survey cycle (E – in reference to the 3.1 CCHS survey from where the sample was drawn and the data linked back)
Position 5:	Variable type (_ - question, D – derived variable)
Positions 6-8:	Question number

For example: The variable from question 103, Smoking Module, BCSS (BCSE_103):

Position 1-2:	BC	Comes from BCSS
Position 3:	S	Smoking Module
Position 4:	E	Cycle 3.1 (callback)
Position 5:	_	underscore (_ = collected data)
Position 6-8:	103	question number & answer option

12.3.2 Variable name component structure in CCHS (as a point of reference)

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

Positions 1-3:	Module/Questionnaire section name
Position 4:	Survey cycle
Position 5:	Variable type
Positions 6-8:	Question number

For example: The variable from question 202, Smoking Module, CCHS Cycle 3.1 (SMKE_202):

Position 1-3:	SMK	smoking module
Position 4:	E	Cycle 3.1
Position 5:	_	underscore (_ = collected data)
Position 6-8:	202	question number & answer option

12.3.3 Positions 1-3: variable / questionnaire section name

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

BCS	British Columbia Smoking History/Lifetime Exposure to Cigarettes
BCM	Mobility Module
BCD	Driving and Safety Module
BCA	Administration Module

12.3.4 Position 4: Cycle and Survey Name

(*note this convention changed for the 2007 CCHS collection, position 4 no longer designates a cycle)

Cycle	Survey Name
A	Cycle 1.1: Canadian Community Health Survey
B	Cycle 1.2: Canadian Community Health Survey, Mental Health and Well-Being
C	Cycle 2.1: Canadian Community Health Survey
D	Cycle 2.2: Canadian Community Health Survey, Nutrition
E	Cycle 3.1: Canadian Community Health Survey
F	BCSS: British Columbia Smoking Survey

12.3.5 Position 5: variable type

_	Collected variable	A variable that appeared directly on the questionnaire
C	Coded variable	A variable coded from one or more collected variables (e.g., SIC, Standard Industrial Classification code)
D	Derived variable	A variable calculated from one or more collected or coded variables, usually calculated during head office processing (e.g., Health 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)

12.3.6 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 a questionnaire becomes simply 01A, and question Q15 becomes simply 15.

For questions which 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” and “no” answer for each response option. For example, if Q2 had 4 response options, the new questions 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.

12.4 Custom tabulations

One way to access the master file is to offer all users the option of having staff in Client Services of the Health Statistics Division prepare custom tabulations. This service is offered on a cost-recovery basis. It allows users who do not possess knowledge of tabulation software products to get custom results. The results are screened for confidentiality and reliability concerns before release. For more information, please contact Client Services at 613-951-1746 or by e-mail at hd-ds@statcan.ca.