The National Population Health Survey – its longitudinal nature

Larry Swain, Gary Catlin and Marie P. Beaudet

Abstract

Objectives

This article discusses some of the benefits and challenges of data from a longitudinal panel as exemplified by the National Population Health Survey (NPHS).

Data source

The NPHS collects both cross-sectional and longitudinal data from a sample of randomly selected individuals. The longitudinal sample will be re-interviewed every 2 years for up to 20 years. Two NPHS cycles have been completed: cycle 1 in 1994/95 and cycle 2 in 1996/97.

Summary

Selected findings from the NPHS are presented to illustrate the benefits of longitudinal data. An overview of questionnaire content, collection methods follows, and sample design is provided. A summary of response rates is followed by a discussion of the methods used to maintain response and to adjust the survey weights in order to reduce nonresponse bias. Confidentiality, dissemination, inconsistencies in reporting, proxy reporting and changes in coding conventions are also discussed.

Key words

health surveys, longitudinal studies

Authors

Larry Swain (613-951-8569; swailar@statcan.ca), Gary Catlin (613-951-8571; catlgar@statcan.ca) and Marie P. Beaudet (613-951-7025; beaumar @statcan.ca) are with the Health Statistics Division at Statistics Canada, Ottawa K1A 0T6.

Some aspects of health have a clear cause-and-effect sequence. Exposure to a virus, for instance, may bring on a bout of the flu. A traffic accident may result in broken bones. However, in most instances, the relationship between cause and outcome is much less obvious. For example, people with higher levels of education and income tend to enjoy better health than do those with less education and lower incomes. But does high socioeconomic status facilitate access to conditions that promote good health, or does good health enable an individual to achieve high socioeconomic status?

Typically, attempts to answer such questions are based on cross-sectional surveys, which gather information about conditions prevailing at one point in time. Inferences about the health consequences of specific characteristics are made by comparing groups with and without the characteristics.

This article is adapted from and expands upon "The National Population Health Survey: Its Longitudinal Nature," by Larry Swain and Gary Catlin, which was presented at the Joint IASS/IAOS Conference, Statistics for Economic and Social Development, September 1998.

But instead of gathering information at various intervals from different people by means of a number of cross-sectional surveys, it would be preferable to look at the same individuals repeatedly and identify changes in their characteristics over time to determine whether there have been corresponding changes in health (panel survey). With such longitudinal data, cause and effect are still difficult to disentangle, but the evidence is stronger because some information on the sequence of events is available. Despite this advantage, previous research and theoretical knowledge must be relied upon to guide the research and the interpretation of findings.

To examine the dynamics of health, Statistics Canada conducts the National Population Health Survey (NPHS). The NPHS collects both cross-sectional and longitudinal data on the physical and mental health of Canadians and their use of health care services. The broad objectives are to:[1]

- measure the health status of the population and its relationship to the use of health care services and determinants of health;
- collect data on the economic, social, demographic, occupational and environmental correlates of health;
- provide information on a panel of individuals who will be followed over time to reflect the dynamic process of health and illness;
- allow the possibility of linking survey data to routinely collected administrative data such as vital statistics, environmental measures, community variables and health services utilization.

This article is an overview of the NPHS content and collection methods. It describes some of the special methodological and operational approaches necessary for the longitudinal and cross-sectional components of the household survey conducted in the 10 provinces.

The first cycle of data collection took place in 1994/95; the second, in 1996/97. The third cycle began in June 1998 and will continue through June 1999. Thus, to date, only two cycles of data are available for analysis. Even so, the results illustrate the unique value of longitudinal information.

In just two years ...

Our health may be influenced by certain physical, social and environmental factors, the effects of which may take many years to emerge. Yet evidence to date, from just two cycles of the NPHS, suggests that some factors have important influences on our health.[2,3]

For instance, in just two years, the harmful effects of smoking are apparent. Among people who had no activity limitations in 1994/95, the odds of becoming limited in daily activities because of respiratory, heart or other circulatory diseases were double for those who smoked regularly, compared with those who had never smoked. This association persisted, even when age, sex, education and household income were taken into account. As well, the odds that smokers who were aged 45 to 84 in 1994/95 would have died by 1996/97 were twice those of people in the same age range who had never smoked.

Seniors generally experience higher rates of chronic disease and loss of function than younger adults. However, this is not necessarily a one-way process. One in five people aged 65 or older who had physical limitations in 1994/95 and who had required help with tasks such as meal preparation or grocery shopping no longer needed such assistance two years later. Improvements in functional ability were also observed among seniors who had had more serious activity limitations and who had relied on others for more basic care such as washing, dressing or eating.

Cross-sectional information may show an apparent stability that is not borne out when longitudinal data are investigated. In 1994/95, for example, an estimated 2.4% of the population aged 18 or older received some type of government-supported home care. The figure was similar in 1996/97, at 2.5%. However, only 36% of people who had received home care in 1994/95 were still receiving services in 1996/97. Almost half of them (46%) were living at home but no longer receiving home care, 14% had died, and 4% had entered institutional health care facilities.

Cross-sectional data reveal a decrease in smoking prevalence from 31% in 1994/95 to 29% in 1996/97

among people aged 15 or older. Longitudinal data provide a look behind the scenes of this overall decline. Among the population who did not smoke in 1994/95, 6% had started smoking for the first time or were relapsed former smokers by 1996/97. In contrast, 14% of those who smoked in 1994/95 had quit by 1996/97.

These early findings illustrate the value of data that can be derived from a longitudinal survey. Such surveys require a number of operational and methodological approaches to maintain, as much as possible, the statistical representativeness of the sample over time and to minimize response error.

Content and collection

In every participating NPHS household, one knowledgeable person provides general demographic, socioeconomic and health information about each household member. This is known as the *general component* (see *Definitions*). In addition, one randomly selected individual—not necessarily the same person—is chosen to provide in-depth information about his or her own health for the *health component*. If the selected respondent is younger than 12, *proxy responses* are obtained. Only the randomly selected individual is followed up every second year for longitudinal purposes.

The questionnaire includes sections of *core* content on health status, use of health services, risk factors, and demographic and socioeconomic characteristics (Appendix Table A). Health status is measured with questions on self-perceived health, functional ability, chronic conditions and activity restriction. The use of health services is measured with questions on visits to health care providers, hospital care and medication use. Behavioural risk factors include smoking, alcohol use and physical activity. Demographic and socioeconomic information includes age, sex, education, ethnicity, income and labour force status.

Each cycle contains *focus* content; that is, additional questions on a specific topic. The focus for the first cycle was psycho-social factors that may influence health, such as stress, self-esteem and mastery. In the second cycle, access to health care services was highlighted, while the third cycle centres on self-care and family medical history. Such focus content is intended for one cycle only, although it may be repeated in future cycles.

Data collection takes place in each of the four seasons, and is done through computer-assisted interviewing (CAI). Cycle 1 was conducted primarily through personal interviews at the selected dwellings. Data for subsequent cycles were and will be collected mainly by telephone. Possible effects of this change in collection methodology have not been researched. All data were obtained by self-reporting (or proxy reporting). No physical measures were taken within the NPHS.

To increase the analytic usefulness of the data, research is under way to link external provincial health records with NPHS responses about use of health care services. Such linkage is done only with the respondent's permission.

Cycle 2—new content

In 1996/97, questions on repetitive strain injuries and alcohol dependence were added to the NPHS health component. The former will appear in all future cycles, while the latter will be added on an occasional basis. As cycle 2 focused on access to health care services, questions were asked about the use of services, reasons for use, barriers encountered, and reasons for non-use or infrequent use. Health services include blood pressure measurements, Pap tests, mammography, breast examinations, physical check-ups, flu shots, dental visits and eye examinations. Questions were also asked about supplementary insurance coverage for dental services, eyeglasses or contact lenses, hospital charges for a private or semi-private room, and prescription medications. A question on the use of emergency services was also included.

To reduce collection costs, health promotion questions sponsored by Health Canada were integrated into appropriate sections of cycle 2. (In 1994/95, a separate supplementary questionnaire had been used.) The topics included sexual health, road safety, actions taken to improve health, and opinions about smoking and alcohol use.

An asthma supplement to cycle 2, sponsored as a separate survey by Health Canada, provided information on the severity of attacks, associated risk factors, management and treatment practices, use of medical services, and asthma education.

As part of an agreement with Alberta Health, supplementary questions were asked in that province about sources of health information, sun exposure, social support, attitudes toward parents, health services, sexual health, violence and personal safety, and coping. Questions on the quality and availability of child health services were asked only in the supplementary samples in Alberta and Manitoba.

Definitions

The *general component* of the National Population Health Survey (NPHS) questionnaire contains demographic, socioeconomic and limited health information about each person in the household. This information is obtained from a knowledgeable person in the household.

The *health component* of the questionnaire contains detailed health questions about one randomly selected individual per household. This information is usually provided by the selected individual.

Proxy responses are those obtained for a particular household member from another member of the household; for example, a parent may answer on behalf of a young child. For the general component of the questionnaire, one person answers on behalf of each member of the household (a combination of non-proxy and proxy responses for households with two or more members). For the health component, the randomly selected individual usually answers on his or her own behalf (non-proxy). However, for children and in some special circumstances, proxy responses are accepted for some topics in the health component. Proxy responses are not permitted for topics such as mental health.

Core content refers to those questions that are asked in each cycle. *Focus* content is special content asked in a specific cycle and integrated into the NPHS questionnaire. *Supplemental* content consists of additional content purchased by a client outside Statistics Canada. This supplemental content is collected in a separate survey or integrated into the NPHS questionnaire.

The *longitudinal (core) sample* consists of the 17,276 randomly selected individuals from the first cycle (1994/95) who will be followed (or recontacted) every 2 years for up to 20 years. To be included, respondents must have completed at least the general component of the questionnaire in 1994/95. Respondents to the 1994/95 supplemental sample were excluded.

Longitudinal data are available from three NPHS files (Table 2). The data in each file are weighted to the 1994 Canadian population.

- The *longitudinal square file* contains data for the 17,276 randomly selected individuals in the longitudinal sample, irrespective of their status in the 1996/97 cycle.
- The *longitudinal full file* includes those for whom complete information is available for both 1994/95 and 1996/97. This file contains data only on the 15,670 individuals from the longitudinal sample who had completed both the general and health components of the questionnaire in 1994/95, and who in 1996/97 had: completed both the general and health components of the questionnaire, were institutionalized and had completed the institutional questionnaire, or had died.
- The *longitudinal partial file* includes those 16,168 individuals for whom at least partial information is available for both 1994/95 and 1996/97—the 15,670 persons on the longitudinal full file as well as the 498 persons who: had completed both the general and health components of the questionnaire in 1994/95, but only the general component in 1996/97; or had completed only the general component of the questionnaire in 1994/95, and in 1996/97 had completed at least the general component, were institutionalized, or had died.

The *cross-sectional sample* for a specific collection cycle consists of the longitudinal sample, the supplemental samples and, where applicable, a top-up sample.

Some provinces purchase s*upplemental samples* (*buy-ins)* for cross-sectional purposes to increase provincial sample sizes for a specific cycle of collection.

A *top-up sample* is added in a specific collection cycle to improve cross-sectional representation of the selected respondents. Without a top-up, the NPHS cross-sectional sample would not adequately represent some segments of the population over time (children and immigrants, for example).

For the purpose of this article, *continuers* are the 15,670 people from the longitudinal full file. *Dropouts* are the 1,606 individuals from the longitudinal sample who either provided partial information in 1994/95 or 1996/97, or were nonrespondents in 1996/97 (that is, they refused to participate or they could not be traced).

Sample design

The NPHS employed a stratified two-stage design[1] (clusters, dwellings) based on Statistics Canada's Labour Force Survey, except in the province of Quebec where Santé Québec's design for the "Enquête sociale et de santé" was used. Base sample sizes for each province were determined using the Kish allocation,[4] which balanced the reliability requirements at national and provincial levels. A minimum of 1,200 households in each province was needed to ensure a specified reliability by sex and broad age groups. Populations on Indian reserves, on Canadian Forces bases, and in some remote areas of Quebec and Ontario were excluded from the household component of the survey. Separate surveys were designed for the North and for health care institutions.

Data were weighted to reflect the sample design, adjustments for nonresponse, and poststratification. For complex survey designs like the NPHS, the usual "textbook" formulas for variance calculations are not appropriate, and more sophisticated methods must be used.[5,6] The jackknife procedure is available to calculate variances for cycle 1; the bootstrap procedure, for cycle 2. These methods take into account the design effect resulting from the complex survey design. Both techniques involve dividing the sample into subgroups (replicates) and calculating a point estimate from a sample composed of a subset of replicates. It is then possible to determine the variation or variance in the estimates from the resulting preset number of samples. Tables of approximate coefficients of variation for both cycles were prepared using these procedures.

Keeping a balanced sample

Because just one member in each sampled household responds to the in-depth health questions and, where applicable, becomes the longitudinal respondent, an individual's chance of being included is inversely related to the number of people in the household. That is, those in smaller households—for the most part single people and the elderly—would be more likely to be selected than would members of larger households (generally parents and dependent children). Although correction could be done with appropriate weighting, this over- or underrepresentation of the longitudinal sample would produce an imbalance of some important analytical domains in meeting prespecified levels of reliability, an imbalance that would continue throughout the 20-year span of the longitudinal panel if not addressed.

To adjust this potential imbalance, a rejective method was applied in cycle 1 to increase the representation of parents and children. A portion of the sample of households, usually between 19% and 40% of the sample, was identified for screening. After the roster of household members was completed, screened households that had no member under age 25 were dropped.[7]

A comparison of the resulting sample with the rejective method and simulated results without the rejective method shows that the approach was quite successful (Table 1). The representation of children (under age 12) and youth (12 to 24) was improved relative to the actual distribution of the population from the 1991 Census. This improvement was not at the expense of parents (generally 25 to 44), whose representation improved as well. Seniors (65 or older) were still overrepresented, but to a lesser degree.

Since this redistribution of the sample was closer to the population distribution, variances decreased for the domains of children and youth and increased for the senior's domain. However, at the overall level, the coefficients of variation increased slightly with the rejective method.

Table 1
Age distribution of randomly selected National Population Health Survey respondents, with and without rejective method, Canada excluding territories, 1994/95

Age group	NPHS respondents: With rejective method	NPHS respondents: Without rejective method	1991 Census
	%	%	%
Total	**100.0**	**100.0**	**100.0**
0- 11	11.9	9.9	16.7
12-24	16.4	13.7	18.2
25-44	33.0	32.2	34.2
45-64	22.1	24.4	20.0
65+	16.5	19.8	11.0

Data sources: *1991 Census; 1994/95 National Population Health Survey*
Note: *Percentages may not add to 100 because of rounding.*

The cycle 2 sample

For cycle 2, a distinction must be made between the longitudinal and cross-sectional samples.

The *longitudinal (or core) sample* contains those randomly selected individuals from the 1994/95 survey for whom at least the general component of the questionnaire had been completed. In total, 17,276 individuals from 1994/95 were eligible for re-interview in 1996/97 (Table 2). This includes 2,022 respondents who were younger than 12 in cycle 1 and who had been interviewed as part of the 1994/95 National Longitudinal Survey of Children and Youth. They were included in the 1996/97 NPHS and will be interviewed in future cycles. To date, for these children, only those data items related to health, socioeconomic or demographic concepts comparable to those collected in the NPHS are contained on the NPHS longitudinal files. For example, chronic conditions and health status are included on the NPHS longitudinal files, but family relationships and parenting practices are not.

Of the 17,276 persons, 16,168 (93.6%) responded in cycle 2. This longitudinal responding sample consists of:

- 15,334 individuals with full information (responses to the general and health components in both cycles);
- 61 who were institutionalized, with a completed institutional questionnaire for the second cycle and full information in the first cycle;
- 275 who had died but for whom full information was provided in the first cycle; and
- 498 with partial information (responses in both cycles but with only the general component in one or both years; some had died or were institutionalized).

The remaining 1,108 were nonrespondents in 1996/97.

Three longitudinal files were created: *full* (15,670 persons = 15,334 + 61 + 275), *partial* (16,168 = 15,670 + 498) and *square* (17,276 persons = 16,168 + 1,108).

Table 2
The National Population Health Survey longitudinal sample, Canada excluding territories, 1994/95 and 1996/97

	Sample size	In 1994/95, responded to: General component	In 1994/95, responded to: Health component	In 1996/97, responded to: General component	In 1996/97, responded to: Health component
Longitudinal square file = 17,276† / Longitudinal partial file = 16,168 / Longitudinal full file = 15,670	15,334	X	X	X	X
	61	X	X	Institutionalized	
	275	X	X	Died	
	498	X	X	X	Nonresponse
		X	Nonresponse	X	Nonresponse
		X	Nonresponse	X	X
		X	Nonresponse	Institutionalized	
		X	Nonresponse	Died	
	1,108	X	Nonresponse	Nonresponse	
		X	X	Nonresponse	

Data source: *National Population Health Survey, 1994/95 and 1996/97*

† Includes 14,786 respondents aged 12 or older with responses to both the general and health components in 1994/95, 468 respondents aged 12 or older with responses to the general component only in 1994/95, and 2,022 respondents who were younger than 12 in cycle 1 and who had been interviewed as part of the 1994/95 National Longitudinal Survey of Children and Youth.

People eligible for the longitudinal sample who moved into an institution, to the Northwest Territories or the Yukon, to an Indian reserve, to a Canadian Forces base, or temporarily out of Canada between the two cycles were followed up.

A total of 2,840 respondents from cycle 1 who were sponsored by the provincial governments that elected to enlarge the sample size in their provinces (buy-ins) were not followed up in 1996/97.

Cross-sectional data collected in cycle 1 (1994/95) are available for 58,439 respondents for the general component of the questionnaire and 17,626 respondents (aged 12 or older) for the in-depth health component (Table 3). For cycle 2 in 1996/97, *supplemental samples (buy-ins)* for cross-sectional purposes in three provinces raised the number of respondents to 210,377 for the general component (173,216 aged 12 or older) and 81,804 for the in-depth health questions (73,402 aged 12 or older). The *cross-sectional sample* for cycle 2 thus consists of the longitudinal sample and the supplemental buy-in samples. There was no *top-up sample* in cycle 2.

The buy-ins for cycle 2 were sponsored by the provincial ministries of health in Alberta, Manitoba and Ontario and were designed to provide pre-specified levels of reliability by health area. These supplemental samples were selected using random digit dialing (RDD) techniques and telephone interviews. Stratification of the RDD samples was based on groups of telephone exchanges. In all three provinces, the general component of the questionnaire was completed for all household members, and one person aged 12 or older was randomly selected for the health component. In Alberta and Manitoba, when possible, a child under age 12 was also selected for the health component.

Response rates in cycle 2

The longitudinal response rate for cycle 2 was 93.6%. For cross-sectional purposes, the response rate for the health component was 93.1% for the longitudinal respondents and 75.8% for the RDD portion among respondents aged 12 or older, for an overall response rate of 79.0%. This excludes the supplementary children selected by RDD in Alberta and Manitoba. For those RDD households with a child younger than 12 and in which an adult had already responded, the response rate for children was 98.2%.

These cross-sectional rates can be broken down into response at the household level and at the selected person level (Table 4). For the longitudinal portion of the sample, the household response rate was 94.3%, with a randomly selected person response rate of 98.7%. For the RDD portion of the sample, the response rate at the household level was 80.0%, and 94.8% for the randomly selected respondent. Overall, the household response rate was 82.6%, and the randomly selected person response rate, 95.6%.

Table 3
The National Population Health Survey cross-sectional samples, Canada excluding territories, 1994/95 and 1996/97

	1994/95			1996/97		
	Total	Longitudinal (core) sample	Supplemental (buy-in) samples	Total	Longitudinal (core) sample	Supplemental (buy-in) samples
General component						
Total	**58,439**	**49,121**	**9,318**	**210,377**	**44,439**	**165,938**
Children (<12)	11,477	9,616	1,861	37,161	8,419	28,742
Adults (12+)	46,962	39,505	7,457	173,216	36,020	137,196
Health component†						
Total	**17,626**	**14,786**	**2,840**	**81,804**	**15,681**	**66,123**
Children (<12)	…	…	…	8,402	1,571	6,831
Adults (12+)	17,626	14,786	2,840	73,402	14,110	59,292

***Data source*:** *1994/95 and 1996/97 National Population Health Survey*

† The 1994/95 NPHS cross-sectional Health file excludes respondents who were younger than 12 and who had been interviewed as part of the 1994/95 National Longitudinal Survey of Children and Youth.

… Not applicable

Confidentiality and dissemination

The National Population Health Survey (NPHS) is conducted under the authority of the Statistics Act, which guarantees that the information remains confidential. All information given to Statistics Canada, whether it is collected through a survey, the Census or any other source, is confidential. The challenge for the NPHS—and other Statistics Canada surveys—is to maximize the data made available while maintaining confidentiality.

For the NPHS, master files are created containing the complete data set. For each cycle, this comprises a cross-sectional General file and a cross-sectional Health file. With cycle 2, a longitudinal file was also produced. Public-use microdata files (PUMFs) are then prepared for use outside Statistics Canada in such a way that individual respondents cannot be identified. These files must be approved by an internal Statistics Canada committee before release. Cross-sectional PUMFs on CD-ROM and diskette have been released for the first two cycles (1994/95 and 1996/97).[8,9] These PUMFs are available for purchase. University researchers and students have access to them through Statistics Canada's Data Liberation Initiative (DLI). The PUMFs are also provided to recipients of the joint Health Canada–Statistics Canada National Health Research and Development Program.

In creating the PUMFs for cycle 1, several steps were taken to meet confidentiality requirements. Univariate counts and combinations of variables were examined at the lowest geographic level in each province to identify unique records in the sample. In addition, the distribution of weights in provinces with supplemental samples was examined to see if particular strata (and therefore, data at a relatively small geographic level) could be identified. The General file was examined to determine if households could be recreated. These procedures led to suppressing or collapsing variables to create the approved NPHS cross-sectional PUMF. In some cases, derived variables were put on the PUMF rather than the more specific responses to individual questions. For example, the derived health utility index instead of the 31 individual questions on which it is based was selected for inclusion on the PUMF.

To produce the 1996/97 cross-sectional PUMFs, a more complex process was required to ensure confidentiality. This was primarily owing to the already released 1994/95 PUMF and the overlap in the cross-sectional samples between the two cycles as longitudinal respondents continued to be surveyed. In addition to the procedures used for 1994/95, the process for 1996/97 included comparing weights between the two cycles to determine if individuals or households could be identified. As well, a more detailed study of three-way tables that examined 13 key variables at the smallest geographic unit by province for which data would be released was carried out to discover unique combinations of variables occurring with high proportions. In addition, to guard against the independent creation by users of longitudinal records from the two cross-sectional PUMFs, the 1994/95 PUMF records were matched to those proposed for 1996/97 based on 12 matching variables to determine the proportion of true (the same individuals) and false (different individuals) matches to the total matches. For the true matches, individual records were reviewed to see if particular variable combinations could lead to identification of individual respondents. New variables in the 1996/97 survey were examined to find out if any threats to confidentiality had arisen since the first cycle. As in 1994/95, these procedures led to the suppression or collapsing of some variables.

Confidentiality concerns also affect the provision of variance estimation tools to users. Since detailed design information necessary for variance estimation techniques would identify geographic areas at a very detailed level, other methods are being examined. However, it is first necessary to ensure that the weights themselves cannot be used to identify small geographic areas or to match cross-sectional PUMFs from different cycles.

The longitudinal aspect of the survey, along with the volume of data, will increase the probability of identifying unique records, and therefore, specific individuals, as more data become available. Given the work required to create the cross-sectional PUMF and the current uncertainty associated with the creation of a longitudinal PUMF after only two cycles of data, it is expected that after the third cycle, verification of confidentiality for the cross-sectional data will be even more complex, and the preparation of a longitudinal PUMF will be impossible.

Because the content of PUMFs is restricted and limited research budgets may preclude users from paying for custom programming, a means of providing broad access to the data must be developed. Some success has been achieved with a service that offers remote access to the master files. Authorized users are provided with dummy NPHS files; that is, files with a similar structure to the master files, but with fictitious data. They prepare their own computer programs, test them on the dummy files and then submit them to Statistics Canada by electronic mail. NPHS staff run the programs on the internal master files, check the output to maintain confidentiality, and send the output to users by electronic mail. For variance estimation, programs can be submitted that use the weights on the master files through remote access. Other means of providing NPHS access to researchers while maintaining confidentiality are being explored (for example, under certain conditions, researchers may have access to the NPHS master files at Statistics Canada's Regional Offices or at Head Office).

Table 4
National Population Health Survey cross-sectional response rates, Canada excluding territories, 1996/97

Level of response	Longitudinal (core) sample	Random digit dialing (RDD) sample	Total (core + RDD)
		%	
Overall response rate†	93.1	75.8	79.0
Household	94.3	80.0	82.6
Randomly selected persons (excluding RDD children)	98.7	94.8	95.6

Data source: *1996/97 National Population Health Survey*
† The overall response rate for randomly selected persons based on all households is calculated by multiplying the response rate for responding households by the response rate for randomly selected persons in responding households.

Maintaining response

Because longitudinal respondents may be in the survey for up to 20 years, maintaining interest and co-operation over this period is essential. The NPHS has several strategies for maintaining and improving response rates. Respondent relations materials such as pamphlets and letters help participants understand the importance and benefits of the survey, and they also address typical concerns such as confidentiality of data (see *Confidentiality and dissemination*).

Given the long-term commitment desired of longitudinal respondents, it was felt that a personal first contact would be beneficial. Therefore, for cycle 1 in 1994/95, an interview was conducted in person at the respondent's dwelling. For subsequent cycles, most interviews were and will be conducted over the telephone, although some respondents may not have telephones or they may prefer a personal interview (about 5% in cycle 2).

Although questionnaire testing is important for any survey, it is especially important for longitudinal surveys. Health experts and focus groups are used to help develop and refine NPHS questions, and the questionnaire is tested twice before each cycle. Such tests help determine if respondents understand what they are being asked. Interviewers provide feedback that is used in the finalization of the questionnaire. Interviewer input also influences the computer screen layouts of the computer-assisted interviewing (CAI) application, which itself undergoes extensive testing.

Because the questionnaire uses a computer application, questions can be personalized for each respondent. This provides acknowledgement of previous contact. For longitudinal respondents, names and sex-specific pronouns and possessive adjectives are integrated into the questions that interviewers read. Past data are embedded in the questionnaire for follow-up. For example, if a longitudinal respondent's highest level of education was determined in an earlier cycle, that information is not requested again. The respondent would be asked only for education obtained since the last interview, with the questions dependent on the level of education already reported.

Tracing is essential for longitudinal surveys. Efforts to locate NPHS respondents who have moved between cycles without notifying Statistics Canada have generally been successful. Only 1.7% of respondents could not be found for cycle 2.

Longitudinal respondents who enter a health care institution or move to northern Canada are contacted, and their data are included in the longitudinal file. Conversely, respondents to the institutional and northern surveys who move to households in the 10 provinces are followed up for their respective longitudinal files.

A "fifth quarter" of data collection was added after cycle 1 for the sole purpose of non-response follow-up.

Nonresponse adjustment and poststratification

Despite efforts to maximize response, some nonresponse is inevitable. This has two possible effects on survey data: a loss in effective sample size and therefore an increase in variance; and biased estimates if nonrespondents differ from respondents in the characteristics measured.[10,11] In fact, for the longitudinal sample, the distribution of cycle 2 *continuers* and *dropouts* differed significantly for sex, age group, work situation, household income, and having one or more chronic conditions

(Appendix Table B). Males accounted for more dropouts than did females: 53.3% compared with 46.7%. More dropouts were from the middle age groups (18 to 24, 25 to 44 and 45 to 64) than from the youngest and oldest groups. Adjustments to the survey weights were applied to compensate for the effects of nonresponse.[12,13]

As a final step in the longitudinal weighting, adjustments were made by province, age group (0 to 11, 12 to 24, 25 to 44, 45 to 64, 65 or older) and sex so that the weighted sample would correspond to the 1994 population estimates (poststratification).

Information available from cycle 1 was used for nonresponse adjustment. The CHAID (Chi-square Automatic Interaction Detection) algorithm within the software Knowledge Seeker was used to form weighting classes for non-response adjustment based on variables considered to be good predictors of nonresponse.[12] Classes were created wherever the greatest differential nonresponse occurred and where there was sufficient sample size for the Chi-squared significance test. Specific variables included income, age, sex, race, place of birth, dwelling owned/rented, presence of children/youths in the household, household size, and several geographic variables such as province and urban/rural designations. The variables used for creation of nonresponse adjustment classes differed from province to province. The adjustments for cycle 2 nonresponse, for the most part, successfully reduced bias among continuers on a subset of NPHS variables (Appendix Table C). Variables used in the nonresponse adjustment, or variables correlated with them, are expected to show the most success in reducing bias among the continuers.

Inconsistencies in reporting over time

The incorporation of past data in the questionnaire design can create special difficulties. Notably, inconsistencies may arise as the same individual is questioned over time. This is compounded for proxy responses.

To minimize such inconsistencies, probes that use data from cycle 1 were built into questions for cycle 2. For example, where change was reported in chronic conditions that typically do not change, verification of the change, reasons for it, and relevant dates were asked. However, if it would have compromised confidentiality, probing was omitted. For example, if a particular chronic condition (such as diabetes or epilepsy) had been recorded previously, but a proxy respondent later reported that the individual did not have that condition, the case would not be probed. Similarly, proxy responses about smoking were not probed. And although probing could validate changes in variables such as alcohol use or weight, it was not done because of the sensitivity associated with these questions.

Data that result from probing were used on the longitudinal file for the current year, but past data were not revised. Inconsistencies were left on the file for analysts to use as deemed appropriate. However, when inconsistencies occurred for date of birth and sex, which were asked in cycle 1 and confirmed in cycle 2, data from cycle 2 were used.

Proxy reporting

Questions on preventive measures and smoking were asked in the health component of the questionnaire, which had a very low proxy response rate (1.8%) for the longitudinal sample aged 12 or older, since proxy reporting was strongly discouraged for this component. In other sections (depression and self-esteem, for example), proxy responses were not permitted. However, questions about chronic conditions and activity restriction were asked in the general component of the questionnaire, where proxy responses may be given for longitudinal respondents.

Proxy reporting decreased in cycle 2 as interviewers tried to contact longitudinal respondents to answer both the general and health components. Nonetheless, inconsistencies suggest that proxy reporting may be a problem. Preliminary comparisons of estimates on the prevalence of chronic conditions from cycle 1 and cycle 2 based on both the general and health components indicate that proxy reporting has an effect, both cross-sectionally and longitudinally. (All data for selected respondents younger than age 12 are based on proxy reporting and were excluded from these analyses.)

Given these preliminary results, the short-term solution for cycle 3 was to improve the interviewer's and procedures manuals to discourage proxy reporting of the general component for longitudinal respondents.

Changes in coding conventions

Another difficulty with longitudinal data occurs when coding conventions change over time; for instance, the classification of diseases, drugs, occupations, industries and geography. The general strategy for the NPHS has been to use the same classification system throughout all cycles. That is, when a new system is adopted, historical data on the longitudinal file will be recoded. Depending on time, space and budgets, more than one classification system could be maintained if this would be meaningful for analysis.

When recoding is not implemented, differences between classification systems are noted in the NPHS documentation.

Concluding remarks

The issues outlined in this article, along with countless others that will no doubt arise during future cycles of the National Population Health Survey, are part of the nature of longitudinal data collection. Such surveys present a challenge to both the researchers who analyze the data, and the survey designers who develop, collect, process and disseminate the information. ●

Acknowledgements

The authors thank Yves Béland, Jiajian Chen, Leslie Geran, Jennifer Hubbard, Bryan Lafrance, Jackey Mayda, Georgia Roberts, Karen Roberts, Eric Sayre, Margot Shields, Kathryn Wilkins and Douglas Yeo for their assistance.

References

1 Tambay J-L, Catlin G. Sample design of the National Population Health Survey. *Health Reports* (Statistics Canada, Catalogue 82-003) 1995; 7(1): 29-38.

2 Statistics Canada. National Population Health Survey: Cycle 2, 1996/97. *The Daily* (Catalogue 11-001) May 29, 1998: 3-5.

3 Statistics Canada. *National Population Health Survey Overview, 1996/97* (Catalogue 82-567) Ottawa: Minister of Industry, 1998.

4 Kish L. Multipurpose sample designs. *Survey Methodology* (Statistics Canada, Catalogue 12-001) 1988; 14(1): 19-32.

5 Rao JNK, Wu CFJ, Yue K. Some recent work on resampling methods for complex surveys. *Survey Methodology* (Statistics Canada, Catalogue 12-001) 1992; 18(2): 209-17.

6 Rust KF, Rao JNK. Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research,* 1996; 5: 283-310.

7 Tambay J-L, Mohl C. Improving sample representativity through the use of a rejective method. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 1995: 29-38.

8 Statistics Canada. *National Population Health Survey, 1994/95: Public Use Microdata Files* (Statistics Canada, Catalogue 82F0001) Ottawa: Minister of Industry, 1995.

9 Statistics Canada. *National Population Health Survey, 1996/97: Public Use Microdata Files* (Catalogue 82M0009XCB, CD-ROM and 82M0009GPE, documentation) Ottawa: Minister of Industry, 1998.

10 Swain L, Dolson D. Current issues in household survey nonresponse at Statistics Canada. *Statistics in Transition*, 1998; 3(3): 439-67.

11 Swain L, Dolson D. Current issues in household survey nonresponse at Statistics Canada. *Nonresponse in Survey Research: Proceedings of the Eighth International Workshop on Household Survey Nonresponse (1997)*, Mannheim: ZUMA, 1998: 1-21. (Note: This is an abbreviated version of reference 10.)

12 Tambay J-L, Schiopu-Kratina I, Mayda J, et al. Treatment of nonresponse in cycle two of the National Population Health Survey. *Survey Methodology* (Statistics Canada, Catalogue 12-001) 1998; 24(2): 147-56.

13 Stukel DM, Mohl CA, Tambay J-L. Weighting for cycle 2 of Statistics Canada's National Population Health Survey. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 1997: 111-6.

14 Rao JNK, Thomas DR. Chi-squared tests for contingency tables. In: Skinner CJ, Holt D, Smith TMF, eds. *Analysis of Complex Surveys*. New York: Wiley, 1989: 89-114.

Appendix

Table A

Summary of content, National Population Health Survey, 1994/95 and 1996/97

	Cycle 1 (1994/95)	Cycle 2 (1996/97)
Core content		
Two-week disability	Y	Y
Health care utilization	Y	Y
Restriction of activities	Y	Y
Chronic conditions	Y	Y
Socio-demographic characteristics	Y	Y
Education	Y	Y
Labour force status	Y	Y
Income	Y	Y
Self-perceived health	Y	Y
Women's health	Y	Y
Blood pressure	Y	Y
Height/Weight	Y	Y
Health status	Y	Y
Physical activities	Y	Y
Repetitive strain injury	N	Y
Injuries	Y	Y
Use of medications	Y	Y
Smoking	Y	Y
Alcohol	Y	Y
Mental health	Y	Y
Social support	Y	Y
Sense of coherence	Y	N
Alcohol dependence	N	Y
Focus content		
Psycho-social		
Stress	Y	N
Ongoing problems	Y	N
Recent negative life events	Y	N
Childhood and adult stressors ("traumas")	Y	N
Work stress	Y	N
Self-esteem	Y	N
Mastery	Y	N
Access to services		
Blood pressure	N	Y
Pap smear test	N	Y
Mammography	N	Y
Breast examination	N	Y
Breast self-examination	N	Y
Breast-feeding	N	Y
Physical check-up	N	Y
Flu shots	N	Y
Dental visits	N	Y
Eye examination	N	Y
Emergency services	N	Y
Insurance coverage	N	Y
Supplemental (buy-in) content integrated into NPHS questionnaire		
Health Promotion Survey		
Diet/Nutrition	N	Y
Height/Weight	N	Y
Breast self-examination	N	Y
Breast-feeding	N	Y
Pregnancy	N	Y
HIV	N	Y
Smoking	N	Y
Alcohol	N	Y
Sexual health	N	Y
Road safety	N	Y
Food insecurity	N	Y
Provincial content (buy-ins)		
Coping (Alberta)	Y	Y
Coping (Manitoba)	Y	N
Health information	N	Y
Tanning and UV exposure	N	Y
Social support	N	Y
Attitudes towards parents	N	Y
Health services	N	Y
Sexual health	N	Y
Violence and personal safety	N	Y
Child health services	N	Y
Supplemental (buy-in) content collected in separate survey		
Health Promotion Survey	Y	N
Asthma Survey	N	Y

Table B
Characteristics at baseline (cycle 1) of continuers and dropouts, longitudinal sample, National Population Health Survey, 1994/95

	Continuers		Dropouts		
Personal characteristics	**Sample size**	**%**	**Sample size**	**%**	**Chi-squared**
Sex					
Male	7,209	49.1	835	53.3	5.3*
Female	8,461	50.9	771	46.7	
Age groups					
0- 11	1,908	16.9	114	11.5	30.6***
12-17	1,047	9.0	90	8.1	
18-24	1,434	8.6	228	12.5	
25-44	5,226	33.3	602	36.1	
45-64	3,509	20.7	378	22.3	
65+	2,546	11.6	194	9.5	
Marital status					
Married	7,416	49.2	806	50.9	ns
Single	5,638	40.5	587	40.0	
Previously married	2,614	10.3	211	9.1	
Work situation					
Currently working	7,288	47.0	810	49.0	18.3***
Worked in past 12 months, but not currently working	1,092	6.2	135	7.6	
Did not work in past 12 months	4,741	25.2	481	28.1	
Not applicable	2,415	21.7	148	15.3	
Household income					
Lowest	1,116	5.3	162	8.4	21.1***
Lower-middle	2,240	12.5	243	15.6	
Middle	4,661	30.5	457	30.4	
Upper-middle	5,192	36.0	439	31.7	
Highest	1,843	15.8	158	14.0	
Educational attainment					
Less than high school	4,744	33.2	518	34.2	ns
High school	1,963	14.8	221	15.2	
Some postsecondary	3,214	23.3	331	20.7	
College diploma or university degree	3,819	28.7	405	29.8	
Live alone					
No	12,749	90.7	1,336	91.6	ns
Yes	2,921	9.3	270	8.4	
Self-reported health status					
Poor	386	1.8	38	2.9	ns
Fair	1,379	7.0	105	7.6	
Good	3,834	23.7	302	25.6	
Very good	5,712	36.0	399	36.5	
Excellent	4,359	31.4	280	27.3	
One or more chronic conditions					
No	7,182	50.1	811	54.2	5.7*
Yes	8,471	49.9	775	45.8	
Suffered depression in last 12 months					
No	12,290	94.5	885	93.0	ns
Yes	781	5.5	68	7.0	

***Data source**: 1994/95 and 1996/97 National Population Health Survey, longitudinal square file*
***Notes**: The distributions exclude the "missing" category. With the exception of household income, self-perceived health and depression, where the percentages missing were 5%, 3%, and 9%, respectively, missing values on the longitudinal square file were less than 1%. The percentages are based on weighted data. The Chi-squared test used 500 bootstrap weights and included the Rao-Scott second-order correction[14] to account for the complex survey design.*
ns Chi-squared value did not reach statistical significance.
** $p \leq 0.05$*
*** $p \leq 0.01$*
**** $p \leq 0.001$*

Table C
Characteristics at baseline of all National Population Health Survey cycle 1 (1994/95) longitudinal respondents and cycle 2 (1996/97) continuers, after and before weight adjustments for cycle 2 nonresponse

	All cycle 1 (1994/95) longitudinal respondents		Cycle 2 (1996/97) continuers only		
			After nonresponse adjustments		Before nonresponse adjustments
Personal characteristics	Sample size	%	Sample size	%	%
Sex					
Male	8,044	49.5	7,209	49.5	49.1
Female	9,232	50.5	8,461	50.5	50.9
Age groups					
0-11	2,022	16.3	1,908	16.3	16.9
12-17	1,137	8.9	1,047	9.0	9.0
18-24	1,662	9.0	1,434	8.9	8.6
25-44	5,828	33.6	5,226	33.6	33.3
45-64	3,887	20.8	3,509	20.8	20.7
65+	2,740	11.4	2,546	11.4	11.6
Marital status					
Married	8,222	49.4	7,416	49.4	49.2
Single	6,225	40.5	5,638	40.4	40.5
Previously married	2,825	10.1	2,614	10.1	10.3
Work situation					
Working	8,098	47.2	7,288	47.6	47.0
Worked in past 12 months, but currently not working	1,227	6.3	1,092	6.3	6.2
Did not work in past 12 months	5,222	25.5	4,741	25.1	25.2
Not applicable	2,563	21.0	2,415	21.1	21.7
Household income					
Lowest	1,278	5.6	1,116	5.4	5.3
Lower-middle	2,483	12.8	2,240	12.5	12.5
Middle	5,118	30.5	4,661	30.3	30.5
Upper-middle	5,631	35.6	5,192	35.9	36.0
Highest	2,001	15.6	1,843	15.9	15.8
Educational attainment					
Less than high school	5,262	33.3	4,744	33.0	33.2
High school	2,184	14.8	1,963	14.8	14.8
Some postsecondary	3,545	23.0	3,214	23.4	23.3
College diploma or university degree	4,224	28.8	3,819	28.8	28.7
Live alone					
No	14,085	90.8	12,749	90.7	90.7
Yes	3,191	9.2	2,921	9.3	9.3
Self-reported health status					
Poor	424	1.9	386	1.8	1.8
Fair	1,484	7.0	1,379	7.0	7.0
Good	4,136	23.8	3,834	23.8	23.7
Very good	6,111	36.1	5,712	36.1	36.0
Excellent	4,639	31.1	4,359	31.3	31.4
One or more chronic conditions					
No	7,993	50.6	7,182	50.2	50.1
Yes	9,246	49.4	8,471	49.8	49.9
Suffered depression in last 12 months					
No	13,175	94.4	12,290	94.5	94.5
Yes	849	5.6	781	5.5	5.5

***Data source**: 1994/95 and 1996/97 National Population Health Survey, longitudinal square and full files*
***Notes:** The distributions exclude the "missing" category. With the exception of household income and depression, where the percentages missing were each 5%, missing values on the longitudinal full file were less than 1%. The percentages are based on weighted data. All weights include the poststratification step.*