

IMPUTATION OF PROXY RESPONDENTS IN THE CANADIAN COMMUNITY HEALTH SURVEY

Martin St-Pierre and Yves Béland¹

ABSTRACT

Between September 2000 and November 2001, the Canadian Community Health Survey (CCHS) collected information on the health of Canadians. The sample contained over 130,000 respondents distributed among 136 health regions in Canada in order to produce reliable estimates at the health region level. Among the respondents, a small percentage were proxy respondents, that is, another person in the household answered on behalf of the selected person. Given the private or personal nature of certain topics in the survey, several questions were not asked by proxy. Therefore, a non-negligible amount of information was missing for these respondents. Considering the magnitude of the situation in some health regions, an imputation of the missing data using a nearest neighbour approach was developed. This article describes the adopted imputation strategy as well as results of simulations done to assess its efficiency.

KEY WORDS: Item nonresponse; Nearest neighbour; Regional data

RÉSUMÉ

Entre septembre 2000 et novembre 2001, l'Enquête sur la santé dans les collectivités canadiennes (ESCC) a recueilli beaucoup d'information sur la santé des Canadiens. L'échantillon contenait plus de 130 000 répondants répartis de façon à produire des estimations fiables pour 136 régions sociosanitaires au Canada. Parmi les répondants, un faible pourcentage étaient des répondants par procuration, c'est-à-dire, qu'une autre personne dans le ménage avait répondu aux questions de l'enquête à la place du répondant choisi. Étant donné le caractère privé ou personnel de certains sujets de l'enquête, plusieurs questions ne pouvaient être demandées par l'intermédiaire d'une autre personne. Par conséquent, une quantité non négligeable d'information était manquante pour ces répondants. Étant donné l'ampleur de la situation dans certaines régions sociosanitaires, une imputation des données manquantes utilisant une approche basée sur le plus proche voisin a été développée. Cet article présente en détails la stratégie d'imputation adoptée ainsi que les résultats des simulations effectuées pour vérifier l'efficacité de celle-ci.

MOTS CLÉS: Données régionales; non-réponse partielle; plus proche voisin

1. INTRODUCTION

In order to address priority health data gaps, Statistics Canada launched the Canadian Community Health Survey (CCHS) in 2000. The main objective of the CCHS is to provide reliable cross-sectional information in order to address data needs at the national, provincial and regional levels. Each cycle of the CCHS consists of two survey components: a health region-level survey the first year with a total sample of more than 130,000 respondents called the *regional component*, and a province-level survey with a sample of 30,000 respondents in the second year referred to as the *provincial component* (Béland, Bailie, Catlin and Singh, 2000). This paper focuses on the regional component of cycle 1, which completed its

collection in November 2001. The primary objective of the regional component is to provide reliable cross-sectional estimates with respect to health determinants, health status and utilisation of the health system for 136 health regions.

During data collection, an unexpectedly high proportion of interviews were completed by proxy. Proxy interviews were allowed only if it was confirmed that the selected respondent would not be present for the entire collection period or if mental or physical incapacity made it impossible for the selected respondent to participate. For a health survey like CCHS, it was expected that this would represent between 3 to 4% of the completed interviews. After six months of collection the proxy rate was 7.6% at the national level. Although field interviewers' procedures

¹ Martin St-Pierre and Yves Béland, Household Survey Methods Division, Statistics Canada, 16th floor, R.H. Coats, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, martin.st-pierre@statcan.ca, yves.beland@statcan.ca.

were re-enforced to rectify the situation, an overall proxy rate of 6.3% was observed at the end of collection. Given the private or personal nature of certain topics in the survey, several questions were not asked by proxy. Consequently, higher rates of item nonresponse were observed. In order to address this specific situation it was decided to impute the missing information caused by the proxy mode of collection. This paper presents the imputation strategy adopted for CCHS. Section 2 gives a short overview of the sample design while section 3 describes the collection operations and the context around the problematic situation. The initial imputation strategy is presented in section 4. Section 5 provides details of simulations that were undertaken to validate the imputation strategy. Section 6 outlines the final strategy implemented in data processing. A conclusion is given in section 7.

2. MAIN ELEMENTS OF THE SAMPLE DESIGN

The CCHS targets persons living in private occupied dwellings who are aged twelve or older. Persons living on Indian Reserves or on Crown lands, residents of institutions, full-time members of the Canadian Armed Forces and residents of certain remote regions are excluded from this survey. The CCHS covers approximately 98% of the Canadian population.

To provide reliable estimates to the 136 health regions (HR) with the survey's budget, a net sample of about 130,000 respondents was needed. The CCHS used three frames to select a sample of households. The majority of the sample of households (~83%) came from the Canadian Labour Force Survey area frame. In some HRs, a Random Digit Dialling (RDD) sampling frame or a list frame of telephone numbers was also used. Approximately 7% of the sample of households came from the RDD frame while the list frame generated almost 10% of the sample (Morano, Lessard and Béland, 2000). The respondents selected from the area frame were interviewed in person while those coming from the phone frames were interviewed by telephone.

The selection of the individual respondents was designed to ensure over-representation of youths (12 to 19) and seniors (65 or older). In approximately 82% of the households selected from the area frame, one person aged 12 or older was randomly selected; two persons (12 or older) were randomly chosen in the remaining households. For households selected from the RDD or the list frames, one person aged 12 or older was randomly chosen from all eligible household members.

The CCHS questionnaire was designed for computer-assisted interviewing. As an important goal of the CCHS

was to collect data on issues of specific relevance to the HRs, the questionnaire was divided into two parts: a common content section of 35 minutes in length containing 30 modules, and a 10-minute optional content section containing modules selected to meet the particular needs of each HR. Provinces and HRs were provided with a choice of 28 questionnaire modules to choose from for their HRs. This resulted in 27 different versions of the questionnaire. The reader is referred to Béland *et al.* (2000) for more details on the sample design.

3. DATA COLLECTION OPERATIONS

The initial plan called for data collection between September 2000 and early October 2001, a period of 13 months. This plan was carefully designed to ensure that the survey's quality objectives would be met. To even out the interviewers' workload and eliminate any seasonal effects, the final sample was randomly divided in 12 so that each month of the year would be properly represented for each HR. A 13th month of collection was planned to provide interviewers with an opportunity for a final attempt to convert non-responding cases.

For most of Statistics Canada's household surveys, collection operations proceed smoothly and within the established parameters. For CCHS, the total workload imposed by the large sample size proved to be a tremendous challenge for the data collection infrastructure in place. To ensure the success of collection operations, a number of established procedures were altered, some more than others. At the end of data collection, a national response rate of 84.7% was achieved (Béland, Dufour and Hamel, 2001).

Proxy reporting. As mentioned earlier, proxy interviews were allowed only if it was confirmed that the selected respondent would not be present for the entire collection period, or in cases of mental or physical incapacity preventing an interview. This would normally represent between 3 to 4% of the completed interviews. Within CCHS, an unexpectedly high proportion of interviews were completed by proxy. At the end of data collection, 6.3% of all interviews were completed by proxy; the rate varied from 2 to 23% at the HR level.

Because of their private or sensitive nature, many CCHS questions or even entire questionnaire modules were appropriate for self-response only, and were skipped when the questionnaire was answered by proxy respondents. Consequently, important information was missing for the individuals represented in those interviews. For CCHS, this represented one third of the questionnaire. Among the common questionnaire modules ten were entirely skipped and two were partially skipped. Among the list of optional

questionnaire modules 21 were skipped. To resolve this item nonresponse an imputation strategy was developed.

4. INITIAL IMPUTATION STRATEGY

4.1 Imputation constraints

Many constraints were taken into account during the development of the imputation strategy. For example, because of the optional content selection aspect of the CCHS which resulted in 27 different questionnaires, only donors who had been administered the same questionnaire could be selected to impute the proxy respondent. In addition, many questions and/or questionnaire modules were only asked of some age-sex groups. There were also a lot of filter questions, i.e., one or more questions were asked only when the respondent had given a certain response to a filter question. To avoid creating inconsistencies in a proxy individual responses, imputation classes, that respect the constraints mentioned above, were created. The selection of a donor was then confined to the same imputation class as the proxy individual.

4.2 Impudon

After evaluation, it was decided to use a Statistics Canada internal SAS macro called IMPUDON (Bissonnette, 2001) to do the imputation.

IMPUDON is a generalized SAS macro for donor imputation. It can find the nearest neighbour using both numerical and non-numerical variables. The non-numerical variables are called *matching variables* or *matching fields*. There are many distance functions available in IMPUDON to find the nearest neighbour. But as most of the CCHS questions to impute were non-numerical i.e. yes/no questions or questions with many response categories it was decided to use the *Total weighted match* distance function for the imputation strategy.

The total weighed match distance function consists of choosing the respondent, from the pool of donor records (non-proxies), with the highest number of weighted matches with the recipient record (proxy). A list of matching fields common to both donor and recipient records was determined (e.g., level of education, type of smoker, etc.) then a weight was assigned to each matching field according to the importance of that field in the imputation model. For matching fields, there was a successful match when the value was the same for both the donor and the recipient records.

4.3 Description of the strategy

Considering the imputation constraints described earlier, it was decided to split the imputation process into three passes. It was felt that it would have been difficult to find donors that satisfy all the conditions to impute all the modules at the same time. It would have created too many imputation classes with only a few units in each class. The three imputation passes, which regroup similar questionnaire modules, are the following:

- 1st pass: Health prevention (11 modules)
- 2nd pass: Mental health (13 modules)
- 3rd pass: Other modules (9 modules).

Proxy respondents were imputed independently from one pass to the other. That is, a recipient record was matched to a different donor record in each pass. Intuitively, it was thought that the quality of the imputation model would be greatly improved if similar questionnaire modules were grouped together. Moreover, relationships between similar questionnaire modules that are imputed during the same pass would be preserved. It should however be noted that these relationships were lost among the modules imputed in different passes.

Having three imputation passes forced the development of an imputation model for each pass. Here are the steps that were followed to perform the imputation for each pass:

- a) Create imputation classes
- b) Identify a list of matching fields
- c) Assign a weight to each matching field
- d) Find the nearest neighbour using the total weighted match distance function.

The first step (a) consisted of creating the imputation classes. Each imputation class constituted the pool of donors to choose from to impute each proxy respondent. For each pass, the imputation classes were defined as a combination of the following factors: different questionnaires (27), sex, age groups and filter questions. The different questionnaires corresponded to either a province or a HR. The age groupings differed between passes and they were formed in such a way that they respected the questionnaire constraints and by taking in account the relationship between the age and the questions to impute. Finally, the imputation classes took into account the filter questions. For example, the module on alcohol dependence which needed to be imputed was only asked to respondents who gave a specific answer to a previous question in the alcohol consumption module.

In order to ensure a sufficient number of donors to choose from, a minimum rule was implemented. The minimum rule was the following:

$$\frac{\text{number of donors}}{\text{number of donors} + \text{number of proxies}} \geq 60\%$$

The next step (b) consisted of choosing the matching fields for each imputation pass. Many characteristics (questions or derived variables) were available for both proxy and non-proxy respondents. The relationships between these characteristics and the questions to impute were studied. The characteristics that had strong relationships (large correlation coefficients) were identified as the matching fields. Of course, the matching fields differed from one imputation pass to the other because of the various nature of the topics between each pass. The list of matching fields was quite extensive but the main ones were the number of visits with a medical doctor, self perceived health, emotional problems index, marital status and chronic conditions.

In step (c), a weight was assigned to each matching field. Initially, a weight of 1 was given to every field. Then that weight was increased to 2, 3 or more if there was a strong correlation with the questions to impute for a given imputation pass.

Taking into account the factors mentioned in (a), (b) and (c), a nearest neighbour was found using the total weighted match distance function described in section 4.2. If there was more than one potential donor (same total of weighted matches) then the donor was randomly selected from them. Here again, a minimum rule was applied in order to perform the imputation. For any recipient record, if the nearest neighbour did not have enough matches then the imputation was not carried out and all fields were left as missing values.

5. SIMULATIONS

In order to assess the efficiency of the imputation strategy, a lot of simulations were undertaken where only the non-proxy respondents for whom responses were available for all the questions were used. Many aspects were evaluated with the simulations. The quality of the imputed values to each question was assessed. Also the usefulness of each matching field and its assigned weight was looked at as well as the size and the efficiency of the imputation classes. Finally the bias and the variance of the estimates after imputation were evaluated.

The following steps describe the strategy implemented to perform the simulations:

- 1) Create a file containing only non-proxy respondents.
- 2) Adjust the sampling weights of the non-proxy respondents using proxy classes to compensate for the proxy individuals dropped from the sample and to have a file that represents the Canadian population.
- 3) Define the imputation model for each imputation pass
 - a) Create the imputation classes
 - b) Choose the matching fields
 - c) Assign a weight to each matching field.
- 4) Imputation simulations
 - a) Generate proxy individuals with partial nonresponse from non-proxy respondents based on a model
 - b) Impute the missing information for the newly generated proxy individuals
 - c) Calculate estimates and success rates of imputation for all the questions by comparing the imputed values with the true values (real answers provided by these individuals).
- 5) Repeat step 3, 100 times.
- 6) Compute results for evaluation
 - a) Overall success rates of imputation
 - b) Biases and variances of the estimates
- 7) Go back to step 2, improve the strategy and re-do steps 3 to 5.

The concept of success rate of the imputed values was used to evaluate the quality of our imputation. The success rate of imputation for a question is defined as the ratio of the number of times the question was imputed correctly (i.e., equal to the true value) over the total number of times the question was imputed. Because the true value was available for each non-proxy respondent, it was possible to compare the imputed value with the true value. At the end of the simulations, success rates for each imputed question were first looked at. Then using the most important imputed questions, overall success rates were calculated by province, by age group and by imputation pass. Also, the overall success rate by imputation pass was calculated depending on the number of matching fields between selected donors and proxy respondents.

Table 1 gives some examples of success rates for some characteristics. The standard deviation gives an idea of the stability of the success rates from one simulation to another. Although success rates were very good for some characteristics they were not as good for other ones. The physical activity (PA) and the self-esteem (SF) indexes had a low success rate while the consultation with a mental health professional in the past 12 months (CM_Q01K) and the suicidal thoughts (SU_Q2) had a high success rate. The other two questions, ever had mammogram (MA_Q030) and the blood pressure check index (BP index), had a fair success rate.

Table 1. Success rates for some questions

Question	Type of Question	Success Rates ¹	Success rates STD ²
CM_Q01K	Yes or No	91.2%	0.3%
SU_Q2	Yes or No	84.8%	0.5%
MA_Q030	Yes or No	68.3%	1.4%
BP index	4 categories	60.0%	0.5%
SF index	3 categories	43.0%	0.9%
PA index	3 categories	42.0%	0.5%

¹ Mean of 100 simulations² Standard deviation of 100 success rates**Table 2. Bias and standard deviation of estimates**

Variable	Geo level	Estimate ¹	Bias	STD ²
CM_Q01K	CAN	8.4%	0.0%	0.06%
= Yes	PROV	6 to 10%	-0.1 to 0.1%	0.07 to 0.21%
PS_Q170	CAN	42.6%	0.0%	0.17%
= Yes	PROV	34 to 49%	-0.4 to 0.1%	0.30 to 1.08%

¹ Mean of 100 simulations² Standard deviation of 100 estimates

As anticipated the imputation models did not introduce bias in the survey estimates. Point estimates were calculated for many characteristics in order to measure the bias introduced by the imputation models. For various domains, estimates were calculated using both true and imputed values and the bias was assessed. In addition to that, small increases on the sampling variances of these characteristics were observed. In general, the increases were larger as the size of the domains of estimation decreased. Table 2 summarizes the results for two typical characteristics: the consultation with a mental health professional in the past 12 months (CM_Q01K) and the PSA blood test (PS_Q170).

6. FINAL IMPUTATION STRATEGY

In light of the results of the simulations and some other issues, the initial imputation strategy was somewhat fine-tuned. The most important change to the strategy was to not impute some questions and/or questionnaire modules. The missing information of those questions and/or questionnaire modules for which the observed success rates were not satisfactory was not imputed; it was left as missing on the data file. It was felt that the impact on the final estimates would have been excessive for some small health regions.

There was also a restructuring of the imputation passes in order to improve the efficiency of the strategy. The

following imputation passes were implemented in the final imputation strategy:

1st pass: Health prevention modules

- 3 modules imputed entirely
- blood pressure, dental visits and eye examinations
- 6 modules imputed partially (some questions only)
- PAP smear test, PSA test, mammography, flu shots, breast examinations, breast self-examinations
- 2 modules not imputed
- physical check-up, smoking cessation aids

2nd pass: Mental health-related modules

- 6 modules imputed entirely
- contact with mental health professionals, alcohol dependence, driving under influence, social support, depression, suicide thoughts and attempts
- 7 modules not imputed
- general health, self-esteem, mastery, spirituality, mood, distress and work stress

3rd pass: Sexual behaviours module

4th pass: Height and weight module (one question)

5th pass: Fruit and vegetable consumption module.

Originally, there were 9 modules to impute in the third pass but it was decided not to impute 6 of the 9 modules (physical activities, sedentary activities, use of protective equipment, changes made to improve health, breastfeeding and patient satisfaction). The remaining three modules were finally imputed in three different passes to better address some specific issues. For the sexual behaviours module (3rd pass), the imputation classes were modified to force the selection of a donor of the same age or no more than five years younger than the respondent to impute. Similarly, in order to improve the imputation of the height and weight module (4th pass), the numerical variable Body Mass Index (BMI) was used to find the nearest neighbour. A list of matching fields was then used only if a donor was not found using the BMI. Finally, the fruit and vegetable consumption module (5th pass) was imputed separately for convenience, as it required its own model.

In the 1st pass, 16 matching fields were used. The sum of the weights for these fields was 30. The imputation was not carried out if there was no donor with at least a weighted match of 23. For the 2nd pass, 17 matching fields were used with a total weight of 22. The minimum rule was set to 15. For the 3rd pass, other than the age, 6 matching fields with equal weight were used to find the nearest neighbour. For the 4th pass, only one matching field was used when the BMI was not sufficient. Finally, 8 matching fields with equal weight were used for the 5th pass.

7. CONCLUSION

After 6 months of data collection of the regional component of the CCHS, an unexpectedly high proportion of interviews (7.6%) were completed by proxy. Consequently, important information was missing for the individuals represented in those interviews because some questionnaire modules were skipped when they were answered by proxy. To alleviate this situation it was decided to impute the missing values of those proxy respondents.

The imputation strategy was developed in a very robust manner. Although not all questions and/or questionnaire modules were imputed it is felt that all steps were undertaken to ensure high efficiency so health analysts can use the data with confidence. The main reason why some questions were not imputed was that there was a lack of good correlated variables available in the CCHS questionnaire.

The diverse nature of the CCHS questionnaire made the imputation exercise a very complex process. Moreover the imputation strategy was developed in such a way that inconsistencies between questions or modules were avoided in order to facilitate data analysis. Numerous analyses and simulations were carried out to set all the criteria for the strategy, such as, deciding which module to impute in which imputation pass, creating the imputation classes, or choosing the matching fields and their weight.

Although the total weighted match distance function to find the nearest neighbour gave good results it is felt that it would benefit from some fine-tuning. It works well with matching fields where the response categories are not a scale (e.g., yes/no, black or white) but there is a loss of information for matching fields where the response categories are a scale (e.g., excellent/very good/good/fair/bad). In this case, the similarity in the

response between donors and the individual to impute is not fully taken in account.

Finally, this imputation process was not planned in advance but was a consequence of the large number of proxy interviews. Considering the extensive amount of work necessary to impute these respondents, it would be preferable to avoid this situation in the future. Therefore, a review of the collection process has already been put in place for the future cycles of the survey and for other surveys in Statistics Canada in order to maintain the number of proxy interviews to an acceptable level.

ACKNOWLEDGEMENTS

The authors would like to thank Johane Dufour for her helpful comments during the development of the imputation strategy.

REFERENCES

- Béland, Y., Bailie, L., Catlin, G. and Singh, M.P. (2000). CCHS and NPHS – An Improved Health Survey Program at Statistics Canada. *Proceedings of the section on Survey Research Methods, 2000*, American Statistical Association.
- Béland, Y. Dufour, J. and Hamel, M. (2001). Preventing Nonresponse in the Canadian Community Health Survey. *Proceedings of Statistics Canada's Symposium, 2001*. Statistics Canada.
- Bissonnette, J. (2001). IMPUDON, Statistics Canada, Internal document.
- Morano, M., Lessard, S. and Béland, Y. (2000). Creation of a dual frame for the Canadian Community Health Survey, 2000 *Proceedings of the Survey Methods Section*, Statistical Society of Canada, pp. 249-254.